ELSEVIER

Signal Processing 77 (1999) 37–48

SIGNAL PROCESSING

www.elsevier.nl/locate/sigpro

Data-driven design and complexity control of time–frequency detectors

Cédric Richard*, Régis Lengellé

Laboratoire LM2S, Université de Technologie de Troyes, 12 rue Marie Curie – BP 2060, 10010 Troyes Cedex, France

Received 15 June 1998

Abstract

In this paper, we introduce a method of designing optimal time–frequency detectors from training samples, which is potentially of great benefit when few a priori information on the nonstationary signal to be detected is available. However, achieving good performance with data-driven detectors requires matching their complexity to the available amount of training samples: receivers with a too large number of adjustable parameters often exhibit a poor generalization performance whereas those with an insufficient complexity cannot learn all the information available in the design set. Then, using the principle of structural risk minimization proposed by Vapnik, we introduce procedures which provide powerful tools for tuning the complexity of generalized linear detectors and improving their performance. Next, these methods are successfully experimented on simulated and real data, with linear detectors operating in the time–frequency domain: it is in such high-dimensional feature spaces that procedures of deriving reduced-bias receivers from training samples are of prime necessity. Finally, we show that our methodology may offer a helpful support for designing detectors in many applications of current interest, such as biomedical engineering and complex systems monitoring. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Artikel stellen wir eine Methode zum Entwurf optimaler Zeit–Frequenz-Detektoren aufgrund von Trainingsdaten vor. Diese Methode ist potentiell von großem Nutzen, wenn wenig A-priori-Information über das zu detektierende instationäre Signal vorhanden ist. Für eine gute Leistungsfähigkeit datengesteuerter Detektoren ist es jedoch erforderlich, deren Komplexität an die Menge verfügbarer Trainingsdaten anzupassen: Empfänger mit zu vielen einstellbaren Parametern besitzen oft eine schlechte Verallgemeinerungsfähigkeit, während solche mit unzureichender Komplexität nicht die gesamte Information erlernen können, die in den für den Entwurf verwendeten Daten enthalten ist. Unter Verwendung des von Vapnik vorgeschlagenen Prinzips der strukturellen Risikominimierung stellen wir weiters leistungsstarke Methoden zur Abstimmung der Komplexität verallgemeinerter linearer Detektoren und zur Verbesserung ihrer Leistungsfähigkeit vor. Diese Methoden werden experimentell anhand simulierter und echter Daten für im Zeit–Frequenz-Bereich arbeitende lineare Detektoren bestätigt: gerade in solchen hochdimensionalen Merkmalsräumen sind Prozeduren, die den datengesteuerten Entwurf von Empfängern mit reduziertem mittlerem Fehler erlauben, von großer Wichtigkeit. Schließlich zeigen wir, daß unsere Methode in zahlreichen Anwendungen von aktuellem Interesse – wie der Biomedizintechnik und der Überwachung komplexer Systeme – eine Unterstützung beim Entwurf von Detektoren bieten kann. © 1999 Elsevier Science B.V. All rights reserved.

* Corresponding author. Tel.: + 33-3-25-71-56-77 (57-82); fax: + 33-3-25-71-56-99; e-mail: cedric.richard@univ-troyes.fr

0165-1684/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0165-1684(99)00021-3

Résumé

Dans cet article, nous exposons dans un premier temps une méthode pour la synthèse de détecteurs temps–fréquence à partir d'un ensemble d'apprentissage, ce qui présente un intérêt majeur lorsque les hypothèses en compétition sont difficilement caractérisables. Il apparaît cependant que les détecteurs résultant d'un processus d'apprentissage présentent un biais important lorsque la quantité de données disponibles est relativement faible. Pour remédier à ce problème, nous présentons dans un second temps plusieurs méthodes basées sur le principle SRM de Vapnik permettant d'adapter la complexité des détecteurs linéaires à la taille de l'ensemble d'apprentissage. Ceci a pour effet d'améliorer significativement leurs performances. Finalement, ces méthodes sont validées à l'aide de données simulées et réelles, pour des détecteurs linéaires opérant dans le domaine temps–fréquence. C'est en effet dans ce type d'espace de représentation de grande dimension que le contrôle de la complexité des détecteurs est primordial, lors du processus d'apprentissage. Enfin, nous montrons que l'approche présentée peut être avantageusement utilisée dans des domaines aussi variés que la surveillance de systèmes complexes ou l'ingénierie biomédicale. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Detectors design; Training method; Complexity control; Time–frequency analysis

1. Introduction

Cohen's class time–frequency (CTFD) representations have been extensively used for detection in applications ranging from radar to machine fault diagnostics, due to the need for dealing with non-stationary signals. Most of the time–frequency (TF) detectors which have been proposed are linear structures operating in the TF domain and are merely equivalent to quadratic receivers usually defined in the time domain [3]. However, a promising TF-based quadratic detection theory has also been introduced in [16–18]: Sayeed et al. identified several scenarios in which detectors are optimum and fully exploit the many degrees of freedom available in the TF representations.

Unfortunately, the design of detectors requires a priori knowledge of signals whereas phenomena are complex in many applications of current interest (e.g., biomedical engineering, complex systems monitoring). Since the collection of labeled signals is often feasible, Jones and Sayeed derived TF detectors directly from training data via the maximization of the Fisher criterion [5]. However, it is stated in [1] that the resulting discriminants can be arbitrarily bad: there are distributions such that even though the two classes are linearly separable, the Fisher linear discriminant has a probability of error close to one. Then, a method of obtaining optimal TF detectors from training samples was developed [12].

It is well known in Pattern Recognition that classifiers designed from training data often have a large bias, particularly when the number of training samples is small compared with the dimension of data [4]. This experimental evidence was theoretically studied by Vapnik and Chervonenkis, who exhibited links between the generalization performance of receivers, their complexity and the size of the training set [20]. Then, the principles of *structural risk minimization* (SRM) [19] and *minimum description length* (MDL) [13] were proposed to match the complexity of classifiers to the available amount of data in order to improve their performance.

In this paper, after a brief description of the usual time and TF linear approaches to decision problems in Section 2, we expose a method of designing optimal linear detectors from training data in Section 3. Then, we show the suboptimality of any linear receiver resulting from the maximization of the Fisher criterion or the signal-to-noise ratio. Finally, we derive a linear detector operating in the TF domain from simulated training data in order to demonstrate the excellent performances of the method. In Section 4, we illustrate the effect of a small training set on the performance of a detector and we briefly justify this phenomenon with a fundamental result of Vapnik and Chervonenkis' theory [20]. Then, we propose methods of improving the performance of linear detectors derived from a small training set and we experiment it on simulated data in the TF domain. Finally, our procedure is successfully applied to a set of EEG events in Section 5. Some conclusions are presented in Section 6.

2. Time and time–frequency approaches to decision problems

2.1. Cohen's class time–frequency representations review

2.1.1. Definition

The Wigner–Ville distribution, which has been extensively studied in recent years, is defined as [2]

$$W_x(t,f) = \int_{-\infty}^{+\infty} R_x(t,\tau)\exp(-2j\pi f\tau)\,d\tau, \tag{1}$$

where $R_x(t,\tau) = x(t+\tau/2)x^*(t-\tau/2)$ is the instantaneous autocorrelation function of the signal x.

This distribution is known for its high resolution in the TF domain and the large number of properties it satisfies [2]. However, its use in practical applications is limited due to the numerous cross-components generated by its bilinear structure. This difficulty can be removed by applying a bi-dimensional filter F to the instantaneous autocorrelation function R_x. This leads us to the following definition of Cohen's class TF distributions (CTFD) [2]:

$$C_x(t,f;F) = \iint_{-\infty}^{+\infty} F(t',\tau)R_x(t+t',\tau) \times \exp(-2j\pi f\tau)\,dt'\,d\tau, \tag{2}$$

where F is called the autocorrelation domain kernel.

2.1.2. Discrete definition

In practice, sampled data are usually processed and F has a finite support S_F defined as

$$S_F = \{(p,m)\in Z : |p| \leqslant P,\ |m| \leqslant M-1\}, \tag{3}$$

in which case the following equivalent of definition (2) can be used:

$$C_x(k,f_i;F) = 2\sum_{m=1-M}^{M-1}\sum_{p=-P}^{P} F(p,m)R_x(k+p,m) \times \exp\left(-\frac{4j\pi m f_i}{2M-1}\right). \tag{4}$$

In the above expression, R_x denotes the discrete instantaneous autocorrelation function of x:

$$R_x(n,m) = x(n+m)x^*(n-m). \tag{5}$$

The properties of discrete CTFD are similar to the continuous time case except for the periodicity in the frequency variable, in which the period is equal to one-half the sampling frequency. To avoid aliasing, this implies that the sampling frequency must be at least twice the Nyquist rate if x is real. Otherwise, the analytic signal $x + jH(x)$ must be used, where $H(\cdot)$ denotes the Hilbert transform: the absence of a negative frequency spectrum eliminates the problem of aliasing which occurs if the signal is sampled at the Nyquist rate [2].

2.2. Time and time–frequency detection frameworks

2.2.1. Linear detection in the time domain

The detection scenario we consider is as follows. Given a discrete-time signal x received over the interval $\{1,\dots,d\}$, where $x = [x(1)\dots x(d)]$, one must decide between the two competing hypotheses H_0 and H_1:

$$\begin{aligned} &H_0:\ x(k) = n(k),\\ &H_1:\ x(k) = s(k) + n(k),\\ &\quad k\in\{1,\dots,d\}, \end{aligned} \tag{6}$$

where s is the underlying signal to be detected and n some additive noise.

The decision between H_0 and H_1 is often made by comparing a test statistic $\lambda(x)$, computed from the observation, to some preset threshold v [8]. As an example, when s is a known deterministic signal and n some white Gaussian noise with a known variance σ^2, it can be shown that the following test statistic is optimal to solve the detection problem (6):

$$\lambda_T(x;s) = \sum_{k=1}^{d} s(k)x(k). \tag{7}$$

This detection structure is called a matched filter. Note that this linear test statistic also maximizes the output signal-to-noise ratio when the noise n is non-Gaussian.

2.2.2. Linear detection in the time–frequency domain

In the perspective of a TF-based detection scheme, the hypothesis testing problem (6) can be rewritten in time and frequency terms using the

WV representation:[1]

$$\begin{aligned} &\mathrm{H}_0\colon\ W_x(k,f_i) = W_n(k,f_i),\\ &\mathrm{H}_1\colon\ W_x(k,f_i) = W_{s+n}(k,f_i),\\ &\quad k\in\{1,\dots,d\},\quad i\in\{1,\dots,d\}. \end{aligned} \tag{8}$$

By analogy with the classical matched filter theory, one can consider the general class of detectors based on linear filtering operations in the TF domain

$$\lambda_{\mathrm{TF}}(x;g_{\mathrm{TF}}) = \sum_{k=1}^{d}\sum_{i=1}^{d} g_{\mathrm{TF}}(k,f_i)W_x(k,f_i), \tag{9}$$

where g_{TF} is a TF reference to be determined using the a priori known statistics of the signal s and the noise n. It is of major importance to note that the key difference between the test statistics (7) and (9) is that λ_{TF} is a quadratic function of the samples $x(k)$ whereas λ_{T} is a linear one. This implies that λ_{TF} can provide an optimal hypothesis test when the signal s to be detected is a random Gaussian signal and n is a white Gaussian noise. It is known that the quadratic test statistic λ_{TF} can equivalently be implemented in the time domain. However, its formulation is much more transparent in the intuitive and physically meaningful TF domain. Moreover, Sayeed and Jones identified several non-stationary composite hypothesis testing scenarios for which TF detectors fully exploit the degrees of freedom available in the TF representations [16].[2]

In Eq. (9), the determination of g_{TF} can be achieved by maximizing the Fisher criterion [5] or the signal-to-noise ratio between the two competing hypotheses [14], when the probability densities of W_x under H_0 and H_1 are unknown. In Section 3, we present an optimum design procedure for linear detectors. Then, this approach is used to optimize g_{TF}.

[1] Any Cohen's class distribution can be used provided that it is invertible to prevent loss of information.

[2] In this paper, we focus on the hypothesis testing problem (8). However, the scope of our methods can easily be extended to the seenarios identified in [16], using the alignment procedure proposed in [15].

3. Optimum design of linear detectors from training data

3.1. Design of linear detectors

3.1.1. Problem formulation and resolution

Linear receivers are optimum (e.g., in the sense of the likelihood ratio) for Gaussian distributions with equal covariance matrices under hypotheses H_0 and H_1. However, even if these assumptions on probability densities are not reasonable in many applications, the simplicity and robustness of this approach often compensate the loss in performance. In this way, we discuss now how to design optimum linear detectors from training data, regardless of the statistics of the observation under H_0 and H_1. This method was introduced by Fukunaga to design linear discriminants in the context of Pattern Recognition [4], and used by Richard and Lengellé to automatically design optimum TF detectors from training data [12].

For a linear test statistic λ, the hypothesis testing problem (6) can be rewritten as

$$\begin{aligned} &\text{if } \lambda(\boldsymbol{X};\boldsymbol{V},v) = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{X} - v \geqslant 0, \text{ then } \mathrm{H}_1\\ &\text{else } \mathrm{H}_0, \end{aligned} \tag{10}$$

where $\boldsymbol{X}$ is the observed signal, $\boldsymbol{V}$ the vector to be determined and v a threshold.

Our design work consists in finding the optimum vector $\boldsymbol{V}$ and threshold value v in the sense of a preselected criterion and for a given data set. Using a minimal a priori knowledge, λ can be characterized by its expected value η_i and variance σ_i^2 under H_i defined as

$$\begin{aligned} \eta_i &= E\{\lambda(\boldsymbol{X};\boldsymbol{V},v)|\mathrm{H}_i\}\\ &= E\{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{X} - v|\mathrm{H}_i\} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{M}_i - v, \end{aligned} \tag{11}$$

$$\begin{aligned} \sigma_i^2 &= \mathrm{Var}\{\lambda(\boldsymbol{X};\boldsymbol{V},v)|\mathrm{H}_i\}\\ &= \boldsymbol{V}^{\mathrm{T}}E\{(\boldsymbol{X}-\boldsymbol{M}_i)(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}|\mathrm{H}_i\}\boldsymbol{V} = \boldsymbol{V}^{\mathrm{T}}\Sigma_i\boldsymbol{V}, \end{aligned} \tag{12}$$

where $\boldsymbol{M}_i = E\{X|\mathrm{H}_i\}$, $\Sigma_i = E\{(\boldsymbol{X}-\boldsymbol{M}_i)(\boldsymbol{X}-\boldsymbol{M}_i)^{\mathrm{T}}|\mathrm{H}_i\}$.

Let Ξ be the class of separability criteria depending only on the parameters η_i and σ_i^2 defined above. Let $\xi\in\Xi$. Since the separability of H_0 and H_1 must be maximized, the derivatives of ξ, with respect to

1. Variables

$\hat{\underline{M}}_i$: sample mean of the observations under H_i

$\hat{\Sigma}_i$: sample covariance matrix under H_i

α : variable ranging between 0 and 1

$\Delta\alpha$: increment

2. Iterative procedure to find α

i) Initializations

- compute $\hat{\underline{M}}_i$ and $\hat{\Sigma}_i$ from the training samples
- set α to 0

ii) While ($\alpha \leq 1$) repeat

- solve $\left[\alpha\hat{\Sigma}_0 + (1-\alpha)\hat{\Sigma}_1\right]\underline{V}_\alpha = \left(\hat{\underline{M}}_1 - \hat{\underline{M}}_0\right)$
- find the threshold v_α which minimizes the estimate of the generalization error E_α
- update α : $\alpha \leftarrow \alpha + \Delta\alpha$

iii) Select the detector $(\underline{V}_\alpha, v_\alpha)$ *which minimizes* E_α

Fig. 1. Iterative algorithm which provides the optimum linear detector in the sense of the best criterion $\xi(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2) \in \Xi$.

η_i and σ_i^2, are equated to zero:

$$\frac{\partial\xi}{\partial \boldsymbol{V}} = \frac{\partial\xi}{\partial\sigma_1^2}\frac{\partial\sigma_1^2}{\partial \boldsymbol{V}} + \frac{\partial\xi}{\partial\sigma_2^2}\frac{\partial\sigma_2^2}{\partial \boldsymbol{V}} + \frac{\partial\xi}{\partial\eta_1}\frac{\partial\eta_1}{\partial \boldsymbol{V}} + \frac{\partial\xi}{\partial\eta_2}\frac{\partial\eta_2}{\partial \boldsymbol{V}} = 0,$$

$$\frac{\partial\xi}{\partial v} = \frac{\partial\xi}{\partial\sigma_1^2}\frac{\partial\sigma_1^2}{\partial v} + \frac{\partial\xi}{\partial\sigma_2^2}\frac{\partial\sigma_2^2}{\partial v} + \frac{\partial\xi}{\partial\eta_1}\frac{\partial\eta_1}{\partial v} + \frac{\partial\xi}{\partial\eta_2}\frac{\partial\eta_2}{\partial v} = 0. \tag{13}$$

In the above equations, the derivatives of η_i and σ_i^2 are given by

$$\frac{\partial\sigma_i^2}{\partial \boldsymbol{V}} = 2\Sigma_i \boldsymbol{V}, \quad \frac{\partial\eta_i}{\partial \boldsymbol{V}} = M_i,$$

$$\frac{\partial\sigma_i^2}{\partial v} = 0 \quad \text{and} \quad \frac{\partial\eta_i}{\partial v} = -1. \tag{14}$$

After the substitution of Eq. (14) into Eq. (13), the resolution of Eq. (13) for $\boldsymbol{V}$ provides a particularly interesting analytical solution for the design of the test statistic:

$$[\alpha\Sigma_0 + (1-\alpha)\Sigma_1]\boldsymbol{V} = (\boldsymbol{M}_1 - \boldsymbol{M}_0),$$
$$\text{where} \quad \alpha = \frac{\partial\xi/\partial\sigma_0^2}{\partial\xi/\partial\sigma_0^2 + \partial\xi/\partial\sigma_1^2}. \tag{15}$$

Thus, the optimum $\boldsymbol{V}_{\text{opt}}$ satisfies Eq. (15) regardless of the selection of ξ since the effect of the criterion only appears in the parameter α ($0 \leqslant \alpha \leqslant 1$). As shown in Fig. 1, we can choose the value of α and the threshold v which minimize the generalization error E_{gene}:

$$\lambda_{\text{opt}} \triangleq (\boldsymbol{V}_{\text{opt}}, v_{\text{opt}}) \triangleq \underset{0\leqslant\alpha\leqslant 1, v}{\text{argmin}} \left(E_{\text{gene}}(\boldsymbol{V}_\alpha, v)\right). \tag{16}$$

Here, $\boldsymbol{V}_\alpha$ satisfies Eq. (15), given α, and E_{gene} is defined as

$$E_{\text{gene}}(\boldsymbol{V}_\alpha, v) = P_0 \int_{\lambda(\boldsymbol{X};\boldsymbol{V}_\alpha, v)>0} p_0(\boldsymbol{X})\,\mathrm{d}\boldsymbol{X} + P_1 \int_{\lambda(\boldsymbol{X};\boldsymbol{V}_\alpha, v)<0} p_1(\boldsymbol{X})\,\mathrm{d}\boldsymbol{X}, \tag{17}$$

where P_i and $p_i(\boldsymbol{X})$ denote the a priori probability and the conditional density of data under H_i, respectively.

It is evident from Eq. (17) that the calculation of the generalization error is, in many practical problems, a difficult task. When we cannot obtain a closed-form expression of E_{gene}, we can compute upper and lower bounds using different techniques such as Leave-One-Out, Resubstitution and Bootstrap methods [4]. The generalization error can also be estimated on a separate test set via an error-counting procedure, i.e., the samples of this set are tested by the detector and the number of misclassified ones is counted. This test error is denoted E_{test}.

As a conclusion, the design procedure presented in this section allows us to determine the optimal linear test statistic λ_{opt} in the sense of the best criterion $\xi \in \Xi$, without setting it up.

3.1.2. Fisher linear discriminant sub-optimality

The Fisher linear discriminant is obtained by maximizing the following criterion [1]:

$$\xi_{\text{Fisher}}(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2) = \frac{(\eta_0 - \eta_1)^2}{P_0\sigma_0^2 + (1-P_0)\sigma_1^2}, \tag{18}$$

where P_i denotes the a priori probability of hypothesis H_i.

The derivatives of $\xi_{\text{Fisher}}(\eta_0,\eta_1,\sigma_0^2,\sigma_1^2)$, with respect to σ_0^2 and σ_1^2, are

$$\frac{\partial \xi_{\text{Fisher}}(\eta_0,\eta_1,\sigma_0^2,\sigma_1^2)}{\partial \sigma_i^2} = -P_i \frac{(\eta_0-\eta_1)^2}{(P_0\sigma_0^2+P_1\sigma_1^2)},$$

$i \in \{0,1\}$.

Therefore, $\alpha = P_0$ and the optimum $\boldsymbol{V}_{\text{Fisher}}$ satisfies

$$[P_0\Sigma_0 + (1-P_0)\Sigma_1]\boldsymbol{V}_{\text{Fisher}} = (\boldsymbol{M}_1 - \boldsymbol{M}_0). \tag{19}$$

In this way, we show that the Fisher linear discriminant is a particular case of Eq. (15), where α is equal to P_0. Consequently, this criterion is not necessarily the best one in the class Ξ. In [1], Devroye et al. stated that Fisher discriminants can be arbitrarily bad: there are distributions such as even though the two classes are linearly separable, the Fisher linear discriminant has a probability of error close to one.

Finally, it should be noted that the signal-to-noise ratio criterion also belongs to the class Ξ: $\boldsymbol{V}_{\text{SNR}}$ satisfies Eq. (15) with α equal to 1.

3.2. Design of linear detectors operating in the TF domain

3.2.1. Principle

The design of the TF detector (9) is straightforward with the algorithm described in Fig. 1. In that case, $\boldsymbol{X}$ and $\boldsymbol{V}$ must be defined as follows:

$$\boldsymbol{X} = [W_x(1,f_1)\, W_x(2,f_1) \ldots W_x(d-1,f_d)\, W_x(d,f_d)]^{\mathrm{T}} \tag{20}$$

and

$$\boldsymbol{V} = [g(1,f_1)\, g(2,f_1) \ldots g(d-1,f_d)\, g(d,f_d)]^{\mathrm{T}}.$$

It should be noted that subspace-based methods must be used to solve Eq. (15) since the matrix $[\alpha\Sigma_0 + (1-\alpha)\Sigma_1]$ is singular, when Σ_i denotes the covariance matrix of (20) under H_i. This result can easily be derived from a property of the discrete WV transform which was recently demonstrated: in [9], it is shown that it only exists $d^2/4 + d/2$ linearly independent TF locations in the $d \times d$ discrete WV representation of a d-sample real signal, and $d^2/4 + d - 1$ in the representation of an analytic one.

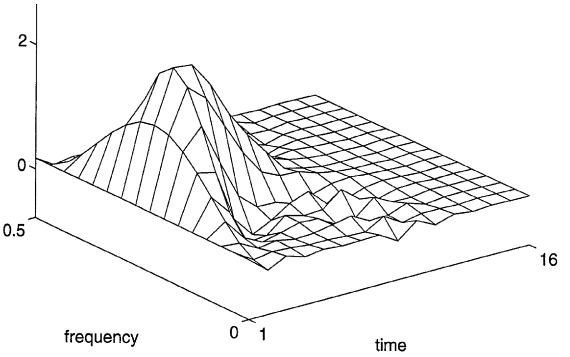

Fig. 2. WV representation of the computer-generated signal s to be detected.

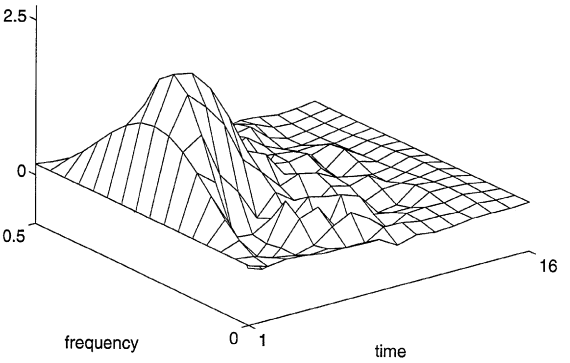

Fig. 3. Data-driven TF detector (training set: 20,000 realizations of H_0 and H_1).

3.2.2. Experiment on simulated data

In the case of detecting the presence or absence of

$$s(k) = k \times \exp(-0.45\,k) \times \sin(0.5\,\pi\,k + \theta),$$

$k \in \{1, \ldots, 16\}$,

in zero mean white Gaussian noise, with phase θ a uniform random variable, the optimal receiver is known to be the inner product of the Wigner–Ville distribution W_s of the signal s to be detected (Fig. 2) with that of the observation x. This detector is called a TF matched filter. In order to illustrate our approach, an experiment of a blind detector design from training data was conducted with 20,000 realizations of hypotheses H_0 and H_1. The TF reference g_{TF} resulting from the training stage is shown in Fig. 3. It appears as an evidence that it has the same structure as W_s.

The performance of this receiver was estimated using 2000 realizations each of signal present or

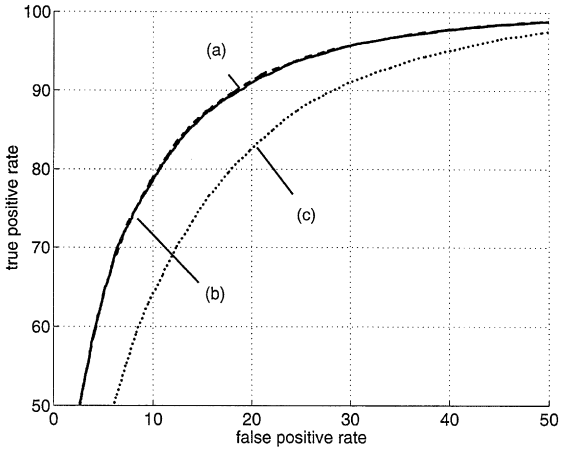

Fig. 4. ROC of (a) the TF matched filter and (b)–(c) the data-driven TF detectors (training sets: 20,000 and 300 realizations of H_0 and H_1, respectively).

absent. Fig. 4 shows the receiver operator characteristics (ROC) of this detector, denoted (b), and the TF matched filter, denoted (a). The latter provides the upper bound on the performance of the various detectors. In this figure, it can be seen that the two ROC are very similar. This illustrates the ability of the proposed method to closely approach the optimal TF detector, when many training samples are available.

However, it is known that receivers designed from training data have a large bias when the number of available samples is relatively small. This experimental evidence was theoretically studied by Vapnik and Chervonenkis, who exhibited links between the generalization performance of receivers, their complexity and the number of design samples [20]. In Section 4, we develop methods of designing reduced-bias detectors. Then, these techniques are used to design reduced-bias TF detectors from training data.

4. Design of reduced-bias linear detectors

4.1. The method of structural risk minimization

Achieving good performances with detectors designed from training samples requires matching their complexity to the amount of available data: receivers with a too large number of adjustable parameters often exhibit poor generalization performances whereas those with an insufficient complexity cannot learn all the information available in the design set. In between, there is an optimal complexity which yields the best generalization error E_{gene} for a given number of training data. This problem is now briefly discussed.

Let χ be a set of detectors, and let VC_χ be its VC-dimension. This parameter characterizes the complexity of receivers which are members of the class χ: it is defined as the maximum number of training samples they can learn without error and for all possible binary labelings. In the case of generalized linear classifiers, for example, this definition implies that VC_χ is given by the number of degrees of freedom available in the structure. Note particularly that VC_χ is bounded by d^2+1 if χ is the class of the linear TF detectors (g_{TF},v) defined in Eq. (9) [9]. However, the determination of VC_χ is generally much more difficult (e.g., if χ is a set of neural networks).

As shown in [19], the VC-dimension of a receiver allows to derive an upper bound of its generalization error from its error rate E_{train} on the training set and the number N of design samples. With probability $1-\eta$, the following inequality holds:

$$E_{\text{gene}} \leqslant E_{\text{train}} + D(N,\text{VC}_\chi,\eta),$$

where

$$D(N,\text{VC}_\chi,\eta) = \sqrt{\frac{\text{VC}_\chi}{N}\left(1+\log\left(\frac{2N}{\text{VC}_\chi}\right)\right)-\frac{1}{N}\log\left(\frac{\eta}{4}\right)}. \qquad (21)$$

The method of *structural risk minimization* introduced by Vapnik et al. [19] consists in matching VC_χ to the amount of training data in order to get the best compromise between the competing terms E_{train} and $D(\cdot)$: reducing VC_χ causes $D(\cdot)$ to decrease but E_{train} to increase. In order to give a precise statement of the VC-dimension selection problem, we assume the following nested sequence of subsets χ_i in the class χ:

$$\chi_1 \subset \cdots \subset \chi_r \subset \cdots \subset \chi$$
$$\text{which implies that } \text{VC}_{\chi_1} \leqslant \cdots \leqslant \text{VC}_{\chi_r} \leqslant \text{VC}_\chi. \qquad (22)$$

Vapnik proposed to approximate the target receiver λ_{opt} in χ as follows:

$$\lambda_{\text{opt}} \triangleq \operatorname{argmin} E_{\text{gene}}\{\lambda \in \chi\}$$
$$\approx \operatorname{argmin} E_{\text{gene}}\{\lambda_{i,\text{opt}} \in \chi_i\}, \quad (23)$$

where $\lambda_{i,\text{opt}} = \operatorname{argmin} E_{\text{train}}\{\lambda \in \chi_i\}$. (24)

When too few data is available to be split into a training and a test set, Vapnik suggested to optimize the upper bound (21) rather than an estimate of E_{gene}.

In the next section, we examine three methods of determining sequences of subsets as in Eq. (22).

4.2. VC-dimension control

4.2.1. Principle of optimal brain damage

One common way of adjusting the VC-dimension of the linear test statistic (10) is to prune some of the components of $\boldsymbol{V}$.[3] This principle is reminiscent of *optimal brain damage* (OBD), a procedure commonly applied after neural networks training [6].

From Eq. (15), we define the best candidate for pruning as the component which minimizes the increase δSE_α of the squared error SE_α defined as

$$\text{SE}_\alpha = \|\Sigma_\alpha \boldsymbol{V} - \boldsymbol{M}\|^2,$$

where

$$\Sigma_\alpha = \alpha\Sigma_0 + (1-\alpha)\Sigma_1 \quad \text{and} \quad \boldsymbol{M} = \boldsymbol{M}_1 - \boldsymbol{M}_0. \quad (25)$$

The increase δSE_α of SE_α can be approximated by

$$\delta\text{SE}_\alpha = \sum_i \frac{\partial \text{SE}_\alpha}{\partial(\boldsymbol{V}^i)}(\delta \boldsymbol{V}^i) + \frac{1}{2}\sum_i \frac{\partial^2 \text{SE}_\alpha}{\partial(\boldsymbol{V}^i)^2}(\delta \boldsymbol{V}^i)^2 + \frac{1}{2}\sum_{i\neq j} \frac{\partial^2 \text{SE}_\alpha}{\partial(\boldsymbol{V}^i)\partial(\boldsymbol{V}^j)}(\delta \boldsymbol{V}^i)(\delta \boldsymbol{V}^j) + \text{O}(\|\boldsymbol{V}\|^2), \quad (26)$$

where $\boldsymbol{V}^i$ denotes the ith component of the vector $\boldsymbol{V}$.[4]

To facilitate the decision about the components $\boldsymbol{V}^i$ to set to zero, the pruning process is performed in a basis of normalized eigenvectors $\boldsymbol{\Phi}_\alpha$ of Σ_α.[5]

In such a basis, SE_α is given by

$$\text{SE}_\alpha = \sum_i (\mu_\alpha^i \tilde{\boldsymbol{V}}^i - \tilde{\boldsymbol{M}}^i)^2,$$

where

$$\tilde{\boldsymbol{V}} = Q^{\text{T}}\boldsymbol{V} \quad \text{and} \quad \tilde{\boldsymbol{M}} = Q^{\text{T}}\boldsymbol{M}. \quad (27)$$

Here, the ith column of the matrix Q is the eigenvector corresponding to the ith eigenvalue μ_α^i of Σ_α. $\tilde{\boldsymbol{V}}^i$ and $\tilde{\boldsymbol{M}}^i$ are the ith components of $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{M}}$, respectively.

If $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}_\alpha \triangleq Q^{\text{T}}\boldsymbol{V}_\alpha$, where $\boldsymbol{V}_\alpha$ satisfies Eq. (15), replacing SE_α in Eq. (26) by (27) yields

$$\delta\text{SE}_\alpha = \sum_i (\mu_\alpha^i)^2 (\delta\tilde{\boldsymbol{V}}_\alpha^i)^2. \quad (28)$$

Consequently, the increase δSE_α^i of SE_α due to pruning the ith component of $\boldsymbol{V}_\alpha$ is as follows:

$$\delta\text{SE}_\alpha^i = [\mu_\alpha^i \tilde{\boldsymbol{V}}_\alpha^i]^2. \quad (29)$$

Thus, the components of $\tilde{\boldsymbol{V}}_\alpha$ corresponding to the smallest increases given by Eq. (29) are good candidates for pruning.

4.2.2. Principle of weight decay (WD) [11]

Vapnik proposed to control the complexity of receivers through an additional penalty term $\gamma\|\boldsymbol{V}_\alpha\|^2$ to be simultaneously minimized with SE_α [19]. In the case of linear detectors, this operation is equivalent to pull the components of $\tilde{\boldsymbol{V}}_\alpha$ to zero predominantly along the principal directions of Σ_α associated with its small eigenvalues since we have

$$\tilde{\boldsymbol{V}}_{\alpha,\gamma}^i \triangleq (\mu_\alpha^i)^2/[(\mu_\alpha^i)^2 + \gamma]\tilde{\boldsymbol{V}}_\alpha^i. \quad (30)$$

Here, μ_α^i and $\tilde{\boldsymbol{V}}_\alpha^i$ are defined as in Section 4.2.1.

As a conclusion, the effect of the penalty term $\gamma\|\boldsymbol{V}_\alpha\|^2$ can be compared to that of a pruning procedure, where γ controls the complexity of $\boldsymbol{V}_\alpha$. By analogy with the OBD procedure, we introduce the following expression to approximate the learning

[3] Let VC_χ^0 be the VC-dimension of the class χ of the linear TF detectors (g_{TF}, v) defined in Eq. (9). $VC_\chi = VC_\chi^\circ - n$ when n components of g_{TF} are proved.

[4] If $\boldsymbol{V}$ satisfies Eq. (15), i.e., $\boldsymbol{V} = \boldsymbol{V}_\alpha$, the first term in Eq. (26) is zero.

[5] The matrix Σ_α can be diagonalized since it is symmetric.

capacity of the receiver $\tilde{V}_{\alpha,\gamma}$:

$$\mathrm{VC}_{\chi_\gamma} = \sum_i \frac{\mu_\alpha^i}{\mu_\alpha^i + \gamma}, \quad \text{given } \alpha. \tag{31}$$

This expression is valid only for broad spectra of eigenvalues [7].

4.2.3. Time–frequency domain partitioning.

In [10], a method of designing reduced-bias linear detectors operating in the TF domain was proposed. It consists in forcing the function g_{TF} in Eq. (9) to be constant over the cells $A_1, \ldots, A_r$ of a partition of the TF domain.

Since the VC-dimension of the class χ_r defined in Eq. (32) is equal to $r+1$, $1 \leqslant r \leqslant d^2$, the principle of *structural risk minimization* described in Section 4.1 can advantageously be used to control the complexity of the TF detector (g_{TF}, v).

$$\chi_r = \{(g_{\mathrm{TF}}, v) | g_{\mathrm{TF}}(k, f_i) = \alpha_p \text{ on } A_p,\ p = 1, \ldots, r\}. \tag{32}$$

Unfortunately, the optimization of the partition $\{A_1, \ldots, A_r\}$ is computationally expensive [10]. As a consequence, this technique is not experimented in the next sections.

4.3. Experiment on simulated data

In order to illustrate the effects of the training set size on the design of a TF detector from training data, the experiment described in Section 3.2.2 was conducted with only 300 realizations of hypotheses H_0 and H_1. The TF reference resulting from the direct application of the training algorithm presented in Section 3 is shown in Fig. 5. It appears as an evidence that the TF component of the signal to be detected is less apparent than in Fig. 3. In addition, Fig. 4 shows that this detector, denoted (c), is the poorest performer of the receivers we synthesized.

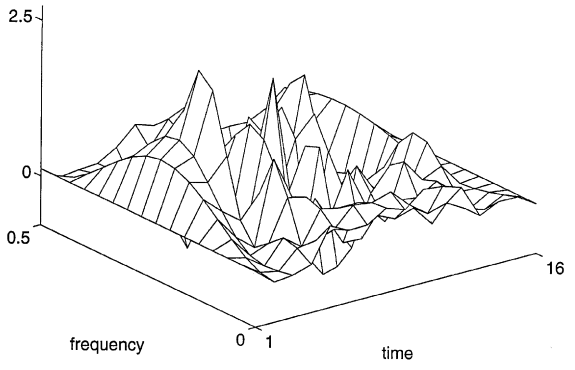

Fig. 5. Data-driven TF detector (training set: 300 realizations of H_0 and H_1).

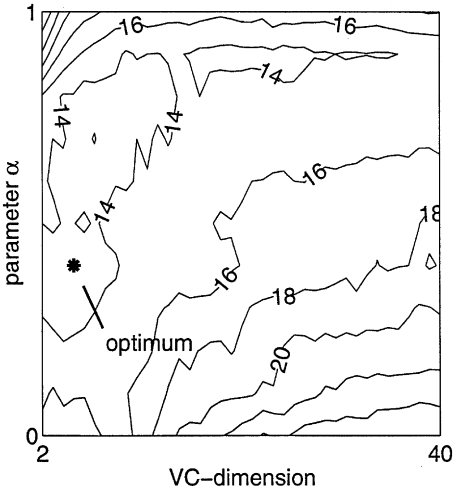

Fig. 6. Error rate of the reduced-bias TF detector as a function of its VC-dimension and α (OBD method).

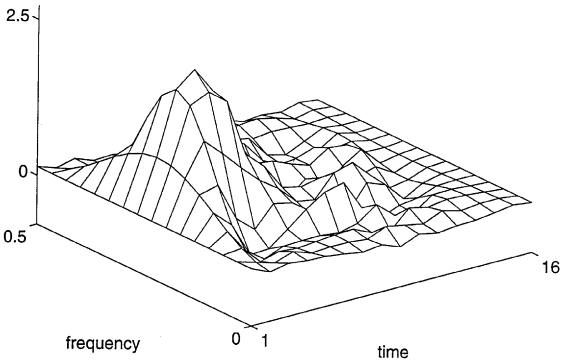

Fig. 7. Reduced-bias TF detector (training set: 200 realizations of H_0 and H_1; OBD method).

For comparison, the design of two reduced-bias TF detectors was performed with 200 realizations of H_0 and H_1, using the OBD and WD procedures. A test set containing 100 other realizations each of signal present or absent was also used to estimate the performance of these receivers during the pruning processes. In the case of the OBD method, the optimization of E_{test}, with respect to the complexity of the detector and the parameter α, resulted in

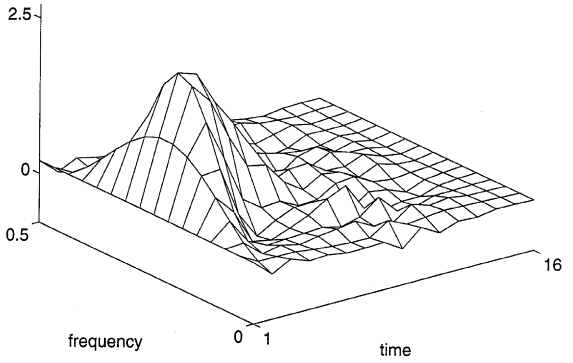

Fig. 8. Reduced-bias TF detector (training set: 200 realizations of H_0 and H_1; WD method).

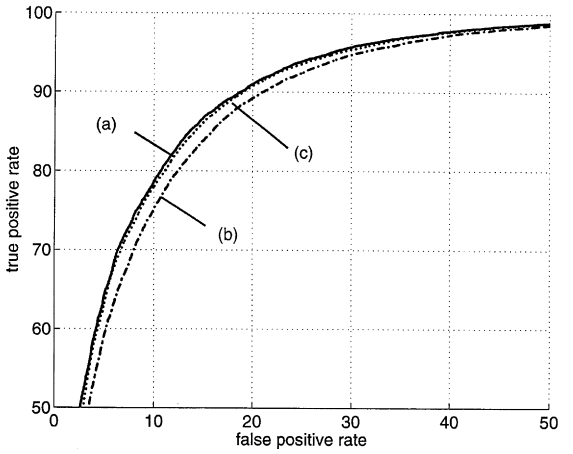

Fig. 9. ROC of (a) the TF matched filter and the (b) OBD and (c) WD-based TF detectors (200 realizations of H_0 and H_1 to train the receiver and 100 to estimate its generalization error during the pruning process).

a minimum for $\alpha = 0.4$ and $VC_\chi = 5$. This result is shown in Fig. 6. It can be seen in Fig. 7 that the reference g_{TF} is quite identical to the WV representation of the signal s. In the case of the WD procedure, E_{test} passed through a minimum for $\alpha = 0.3$ and $\gamma \approx \mu_{\alpha=0.3}^{i=3}$. The resulting reference g_{TF}, which is also similar to W_s, is shown in Fig. 8.

Finally, the generalization performance of these detectors was estimated with 2000 realizations each of signal present or absent, as shown in Fig. 9. We supposed that $P_0 = P_1 = 1/2$. Using the TF matched filter, which is the optimal detector, the generalization error was equal to 14.11%. The generalization error of the TF receiver which was trained with only 300 realizations of H_0 and H_1 was equal to 18.49%. This result must be compared to 15.45% and 14.37% obtained with the OBD and WD based TF detectors, respectively. This clearly demonstrates the ability of the proposed methods to approach the performance of the optimal detector, even if the size of the training set is relatively small compared to the dimension of the problem.

5. Experiment on real data

5.1. The detection of K-complexes in sleep EEG

Automated detection of waveforms such as alpha, delta and K-complex waves in the EEG is an important component of sleep stage monitoring. The K-complex is one of the key features that contributes to sleep stages assessment. This transient EEG pattern has a total duration of between 500 and 1500 ms and is roughly characterized by a sharp upward wave followed by a downward one. Its amplitude is three times background activity and is generally larger than 75 μV. The automated detection of the K-complex is difficult due to the stochastic nature of the EEG: the K-complex can

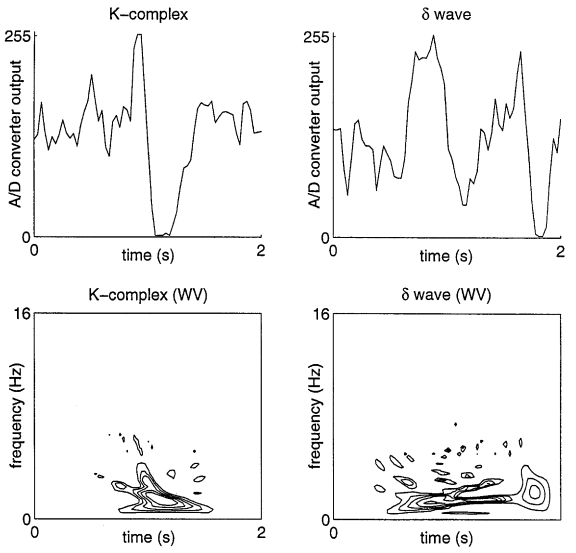

Fig. 10. Examples of EEG events in the time and the TF domains.

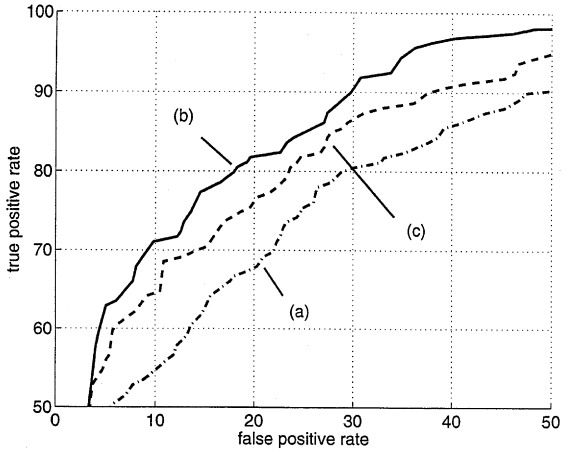

Fig. 11. Detection of K-complexes in sleep EEG. ROC of (a) the TF detector, (b) the reduced-bias TF detector and (c) the reduced-bias T detector.

have a large variety of shapes and it is not always distinctly different from the EEG background activity, as shown in Fig. 10.

The EEG signals used for the design of the detector were stored from the channel Cz. The raw EEG data was digitized with an 8 bit A/D converter at a sampling frequency of 128 Hz and segmented as follows. Two-second intervals, either containing K-complexes or paroxysmal delta bursts bearing some resemblance to K-complexes, were selected. Then, each segment was cropped by placing the x-axis intersection, present between the upward and downward peaks, in the middle of the interval. Each segment was also decimated by a factor of 8 before the evaluation of its Wigner–Ville representation. This data was further split into a training set (300 K-complexes and 600 delta waves) that was used to design detectors of various classes and a test set (150 K-complexes and 300 delta waves) from which their generalization error was estimated during the pruning process. The remaining signals (159 K-complexes and 296 delta waves) were used to asses their performance via empirical ROC. The results are presented in Fig. 11 and discussed below.

5.2. Discussion

(i) *TF detector*. The rank of the 1024×1024 matrices Σ_α, calculated from the WV distribution of our 32-sample analytic signals, was equal to 287. This conforms to the theory evoked in Section 3.2.1. Although we then use a subspace based method to solve Eq. (15), the resulting TF detector, denoted (a) in Fig. 11, performed poorly. This might be partly due to the insufficient number of training data compared with the dimension of the problem.

(ii) *Reduced-bias TF detector*. The design of an OBD-based TF detector was performed. Its error rate E_{test} on the test set passed through a minimum for $\alpha = 0.6$ and $\text{VC}_\chi = 54$. This resulted in improved performances of the reduced-bias TF detector, denoted (b) in Fig. 11, as compared to (a).

(iii) *Reduced-bias T detector*. For the sake of comparison, the linear statistic (7) operating in the time domain was designed and its VC-dimension was optimized using the OBD procedure. As shown in Fig. 11, this receiver (c) performed better than (a) but poorer than (b). This justifies the use of a quadratic test statistic rather than a linear one to solve this problem of detection. However, a pruning procedure must be used to improve the performance of the quadratic receiver since few training data are available.

This experiment demonstrates the ability of our methodology to design a reduced-bias detector of K-complexes without prior knowledge of these events. This is of great benefit in this application since phenomena are so complex and poorly understood that there is little hope of well modeling.

6. Conclusion

In this paper, we have introduced a method of designing optimal generalized linear detectors which requires no prior knowledge of the event to be detected. These receivers, which are directly derived from training data, theoretically perform better than those obtained via the maximization of the Fisher criterion and the signal-to-noise ratio.

However, it is well known in Pattern Recognition that the generalization error of classifiers strongly depends on their complexity and the number of available training samples. The procedures developed here, which are based on the principle

of *structural risk minimization* developed by Vapnik, provide powerful tools for tuning the VC-dimension of linear detectors of various kinds and improving their performance.

Finally, we have successfully experimented our approach on simulated and real data, with linear detectors operating in the time and time–frequency domains. In particular, an experiment on a set of EEG events has pointed out its ability to design a reduced-bias TF receiver without prior knowledge of phenomena.

As a conclusion, this blind methodology may offer an helpful support for designing efficient detectors in many applications of current interest, such as complex systems monitoring and biomedical engineering.

References

[1] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, NY, 1996.
[2] P. Flandrin, Temps–Fréquence, HERMES, Paris, F, 1993.
[3] P. Flandrin, A time–frequency formulation of optimum detection, IEEE Trans. Acoustics Speech Signal Process. 36 (9) (1988) 1377–1384.
[4] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, London, 1990.
[5] D.L. Jones, A.M. Sayeed, Blind quadratic and time–frequency based detectors from training data, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, pp. 1033–1036.
[6] Y. Le Cun, J. Denker, S. Solla, Optimal brain damage, in: D.S. Touretzky (Ed.), Advances in Neural Information Processing 2, Morgan Kaufmann, San Mateo, 1989, pp. 598–605.
[7] J.E. Moody, Generalization, weight decay and architecture selection for non-linear learning systems, in: J.E. Moody, S.J. Hanson, R.P. Lippmann (Eds.), Advances in Neural Information Processing, Vol. 4, Morgan Kaufmann, San Mateo, 1992.
[8] H.V. Poor, An Introduction to Signal Detection in Noise, Springer, New York, NY, 1994.
[9] C. Richard, R. Lengellé, On the dimension of the discrete Wigner–Ville transform range space. Application to time–frequency detectors design, in: Proc. IEEE-SP International Symposium on Time–Frequency and Time-Scale Analysis, Pittsburgh, PA, 1998, pp. 5–8.
[10] C. Richard, R. Lengellé, Structural risk minimization for reduced-bias time–frequency-based detectors design, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, 1998, pp. 2397–2400.
[11] C. Richard, R. Lengellé, Two algorithms for designing optimal reduced-bias data-driven time–frequency detectors, in: Proc. IEEE-SP International Symposium on Time–Frequency and Time-Scale Analysis, Pittsburgh, PA, 1998, pp. 601–604.
[12] C. Richard, R. Lengellé, Une nouvelle approche pour la détection linéaire optimale dans le plan temps–fréquence, in: Proceedings of the Seizieme Colloque GRETSI, Grenoble, F, 1997, pp. 659–662.
[13] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific, Teaneck, 1989.
[14] B. Samimy, G. Rizzoni, A.M. Sayeed, D.L. Jones, Design of training data-based quadratic detectors with application to mechanical systems, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1996, pp. 1767–1770.
[15] A.M. Sayeed, Data-driven time–frequency and time-scale detectors, in: Proceedings of the SPIE's 42nd Meeting, San Diego, 1997.
[16] A.M. Sayeed, D.L. Jones, Optimal detection using bilinear time–frequency and time-scale representations, IEEE Trans. Signal Process. 43 (12) (1995) 2872–2883.
[17] A.M. Sayeed, D.L. Jones, Optimal reduced-rank time–frequency/time-scale quadratic detectors, in: Proceedings of the IEEE-SP International Symposium on Time–Frequency and Time-Scale Analysis, Paris, F, 1996, pp. 209–212.
[18] A.M. Sayeed, D.L. Jones, Time–frequency detectors, in: Proceedings of the CISS'96, Princeton, NJ, 1996.
[19] V.N. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, NY, 1982.
[20] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.