Signal Processing 18 (1989) 349-369
Elsevier Science Publishers B.V.

DISTANCE MEASURES FOR SIGNAL PROCESSING AND PATTERN RECOGNITION*

Michèle BASSEVILLE

IRISA/CNRS, Campus de Beaulieu, F-35042 Rennes Cedex, France

Received 2 August 1988
Revised 23 February 1989 and 26 June 1989

Abstract. Some general tools for measuring distances either between two statistical models or between a parametric model (or signature) and a signal are presented. These tools are useful in a variety of signal processing applications such as detection, segmentation, classification, recognition and coding.

After a section devoted to general distance measures between probability laws, the question of spectral distances between processes is investigated. Then results concerning AR and ARMA models are described. Problems related to the interaction between distances for parametric models and estimation of the parameters of these models are also mentioned. Also recalled (when necessary) are some classical results about error bounds in classification and feature selection for pattern recognition, which are obtained with the aid of properties of distance measures.

Zusammenfassung. Wir stellen einige allgemeine Werkzeuge zur Messung von Abständen zwischen zwei statistischen Modellen oder zwischen einem parametrischen Modell und einem Signal vor. Diese Werkzeuge sind hilfreich bei der Lösung einer Reihe von Signalverabeitungs-Aufgaben wie etwa der Detektion, der Segmentierung, der Klassifizierung, der Erkennung oder der Kodierung.

Nach einem Abschnitt, der allgemeinen Abstandsmaßen zwischen Wahrscheinlichkeits-Gesetzen gewidmet ist, untersuchen wir die Frage spektraler Abstände zwischen Prozessen. Dann beschreiben wir Ergebnisse, die AR- und ARMA-Modelle betreffen; hierbei gehen wir auch auf das Problem der gegenseitigen Beeinflussung von Abstandsmessung für parametrische Modelle und Parameter-Schätzung für diese Modelle ein. Weiterhin erinnern wir (wo nötig) an einige klassische Ergebnisse bezüglich der Fehlergrenzen und der Merkmals-Auswahl bei der Mustererkennung, die man mit Hilfe der Eigenschaften von Abstandsmaßen gewinnen kann.

Résumé. On se propose de présenter quelques outils généraux pour mesurer des distances soit entre deux modèles statistiques soit entre un modèle paramétrique et un signal. Ces outils sont utiles pour résoudre de nombreux problèmes en traitement du signal et notamment pour la détection, la segmentation, la classification, la reconnaissance ou le codage.

Après un paragraphe consacré à des mesures générales de distances entre lois de probabilité, on considère le problème des distances spectrales entre processus. Puis on présente des résultats relatifs aux modèles AR ou ARMA, pour lesquels on mentionne aussi les problèmes liés à l'interaction entre distances de modèles paramétriques et estimation des paramètres de ces modèles. Sont également rappelés, lorsqu'il y a lieu, les résultats classiques concernant les bornes d'erreur de classification ou la sélection de traits caractéristiques pour la reconnaissance des formes, résultats obtenus à l'aide de propriétés de distances précisément.

Keywords. Distances, detection, classification, segmentation, recognition, coding.

1. Introduction

Distance measures between statistical models or between a model and observations are widely used concepts in signal processing (and in automatic control) for solving various problems such as detection, automatic segmentation, classification, pattern recognition, coding, (model validation,

* This work was supported by C.N.R.S.-Groupement de Recherche no. G0134 "Traitement du Signal et Image".

choice of optimal input signals for system identification) ....

To our knowledge, the studies concerning distance measures are basically of two types, apart from those of probabilists and statisticians. On one hand, there are general studies for the computation of error probabilities in classification problems (of any objects characterized by any measurements), without taking into account either the nature of the parameters which characterize the probability laws or the way by which they have been estimated. On the other hand, there are many specific studies in the speech processing domain (coding, recognition), where refinements of Itakura or cepstral distance measures still emerge.

The aim of this paper is to gather scattered tools and results concerning distance measures, in view of applications in signal processing, for detection and recognition in general. Especially, we shall address some typical issues in model based signal processing, namely the choice of models, parametrizations and parametric estimation methods on one hand, and the choice of distance measures between these models on the other hand, without forgetting the possible interaction between these two choices. However we do not claim to have exhaustively compiled all the literature concerning distance measures. Nevertheless, we try to follow a presentation going from a general framework to particular cases.

In Section 2, we introduce general distance measures between probability laws and the relationships existing among them. Then, we present some general tools for measuring the distance between a model and a signal. Section 3 is devoted to spectral distance measures between processes. In Section 4, we analyse the results related to AR or ARMA models.

Let us emphasize that the word distance here means a measure of how far away from each other the laws are, and is not used with the strict sense it has in metric spaces. Particularly, the measures which are mentioned are not all symmetrical and do not all satisfy the triangular inequality. Furthermore, we could have increased the homogeneity among the various distance measures by constraining all of them to be homogeneous either to quadratic quantities or to square roots of quadratic quantities (see (6), (29) and (33) for example). We did not do so, thus preserving the original definitions of these distance measures.

Furthermore, we shall use throughout the paper the following terminology and notations:

- $d(P_1, P_2)$: distance between the probability laws P_1 and P_2.
- $d(A_1, A_2)$: distance between the parametric models A_1 and A_2.
- $d(y_1, A_2)$: distance between a signal y_1 and a model A_2.

With such notations, if $\hat{A}_i$ denotes an estimate of A_i, $i = 1, 2$, then

$$d(\hat{A}_1, \hat{A}_2) \quad \text{and} \quad d(y_1, \hat{A}_2)$$

are distances between signals. The symbol $\hat{}$ will often be omitted for simplification.

2. General distance measures

In this section, we introduce general classes of distance measures, or divergence coefficients, between probability distributions. In Section 2.1, we start with the class related to Csiszar f-divergence [10]. This class contains many known distance measures which we also recall in the multidimensional case. The general $\bar{\rho}$ distance, either between random variables or between random processes, is introduced in Section 2.2 [21]. Another general class of dissimilarity measures, based on the so called Jensen difference, is presented in Section 2.3 [47] together with a general tool for associating, to each divergence measure between parametric laws, a metric on the parameter space. Then in Section 2.4 we describe the so called class of general mean distance introduced by Boekee and Van der Lubbe [6] for pattern recognition. In Section 2.5 we investigate a general contrast criterion which may be used as a distance and which was introduced by Poor [45] for robust detection. Then we describe some general tools for

measuring the distance between a model and observations: In Section 2.6, we recall the axiomatic derivation of the entropy principle due to Shore [48], and finally in Section 2.7 we present a general model validation tool to be used for segmentation or monitoring [4].

2.1. f-divergence

This general notion seems to have been introduced by Csiszar [10-12] and independently by Ali and Silvey [1]. It is based upon the fact that it is intuitively "natural" to measure the distance between two probability distributions p_1 and p_2 with the aid of the "dispersion"—with respect to p_1—of the likelihood ratio.

2.1.1. Definition

More precisely, let λ be a measure on a space $(\mathscr{X}, \mathscr{F})$ such that any probability laws P_1 and P_2 are absolutely continuous with respect to λ, with densities p_1 and p_2 (e.g. $\lambda = P_1 + P_2$ or Lebesgue measure). Let f be a continuous convex real function on $\mathbb{R}_+$, and let g be an increasing function on $\mathbb{R}$. Consider the following class of divergence coefficients between two probabilities:

$$d(P_1, P_2) = g\left[\mathbb{E}_1\left[f\left(\frac{p_2}{p_1}\right)\right]\right], \tag{1}$$

where

$$\frac{p_2}{p_1} \triangleq \phi \tag{2}$$

is the likelihood ratio, and where $\mathbb{E}_1$ is the expectation with respect to P_1.

2.1.2. Properties

Then d has the following properties [1]:

(1) *If $y = t(x)$ is a measurable transformation from $(\mathscr{X}, \mathscr{F})$ to $(\mathscr{Y}, \mathscr{G})$ then*

$$d(P_1, P_2) \geq d(P_1 t^{-1}, P_2 t^{-1}), \tag{3}$$

where $P_i t^{-1}$ is the measure image of P_i by t.

This implies that, when t is the selection of coordinates of a process $(x_n)_{n\in\mathbb{N}}$, we do not decrease the distinguishability between the two laws when we increase the number of observations, i.e.,

$$d(P_1^{(m)}, P_2^{(m)}) \leq d(P_1^{(n)}, P_2^{(n)}), \quad \text{for } m < n, \tag{3a}$$

where the $P_i^{(j)}$ are the marginal laws of $x_1, \ldots, x_j$.

(2) *$d(P_1, P_2)$ is minimum when $P_1 = P_2$ and maximum when $P_1 \perp P_2$.*

(3) *If $(p_\theta;\ \theta \in]a, b[)$ is a family of densities on $\mathbb{R}$ with monotone likelihood ratio* [37, p. 68], *then for $a < \theta_1 < \theta_2 < \theta_3 < b$, we have*

$$d(P_{\theta_1}, P_{\theta_2}) \leq d(P_{\theta_1}, P_{\theta_3}). \tag{4}$$

Let us notice that the convexity of f is a necessary condition for Property 1. Furthermore, for g identity and p_1, p_2 two densities, we have [1]:

$$d(p_1, p_2) = \int_{\mathscr{X}} f\left(\frac{p_2(x)}{p_1(x)}\right) p_1(x)\, dx \geq f(1),$$

with equality if and only if $p_1 = p_2$ almost everywhere.

A key issue here is that there exist [1] other measures of the dispersion of ϕ which are not the expectation of a convex function of ϕ. Thus it is possible to build divergence coefficients (or distance measures) based upon ϕ which do not have form (1), and of course coefficients which are not based upon ϕ (see Section 2.2). However, we shall see that (1) contains many usual measures, and thus the comparison between many distance measures reduces to the comparison between convex functions [6]. Furthermore, the classification error probability P_e, for which the search for upper and lower bounds—see (18)-(21)—gave rise to many studies about distance measures [6, 8, 15], can also be written as in (1) with $f(x) = -\min(x, 1-x)$. Thus the search for upper and lower bounds for P_e reduces to compare this function f to other convex functions [5]. Other divergence measures between multinomial distributions are studied in [7] together with their convexity properties. See also [46].

2.1.3. Examples

- *Variational distance*

$$f(x)=|1-x|, \qquad g(x)=\tfrac{1}{2}x,$$

$$d(P_1, P_2)=\tfrac{1}{2}\int_{\mathscr{X}} |p_2-p_1|\,\mathrm{d}\lambda \triangleq V(P_1, P_2). \tag{5}$$

- *Hellinger distance*

$$f(x)=(\sqrt{x}-1)^2, \qquad g(x)=\tfrac{1}{2}x,$$

$$d(P_1, P_2)=\tfrac{1}{2}\int_{\mathscr{X}} (\sqrt{p_2}-\sqrt{p_1})^2\,\mathrm{d}\lambda \triangleq H^2(P_1, P_2) \tag{6}$$

- *Kullback information*

$$f(x)=-\mathrm{Log}\, x, \qquad g(x)=x,$$

$$d(P_1, P_2)=\int_{\mathscr{X}} p_1 \mathrm{Log}\frac{p_1}{p_2}\,\mathrm{d}\lambda \triangleq K(P_1, P_2). \tag{7}$$

- *Kullback divergence*

$$f(x)=(x-1)\,\mathrm{Log}\, x, \qquad g(x)=x,$$

$$d(P_1, P_2)=\int_{\mathscr{X}} (p_2-p_1)\,\mathrm{Log}\frac{p_2}{p_1}\,\mathrm{d}\lambda = J(P_1, P_2) \triangleq K(P_1, P_2)+K(P_2, P_1) \tag{8}$$

which is symmetrical.

- *Chernoff distance*

$$0\leqslant r\leqslant 1; f(x)=-x^{1-r}, \qquad g(x)=-\mathrm{Log}\,(-x), \qquad d(P_1, P_2)=-\mathrm{Log}\, C(P_1, P_2),$$

where

$$C(P_1, P_2)=\int_{\mathscr{X}} p_1^r p_2^{1-r}\,\mathrm{d}\lambda, \tag{9}$$

is called Chernoff coefficient or Hellinger path.

- *Bhattacharyya distance*: Previous case with $r=\frac{1}{2}$, i.e.,

$$f(x)=-\sqrt{x}, \qquad g(x)=-\mathrm{Log}\,(-x), \qquad d(P_1, P_2)=-\mathrm{Log}\,\rho(P_1, P_2) \triangleq B(P_1, P_2),$$

where

$$\rho(P_1, P_2)=\int_{\mathscr{X}} \sqrt{p_1 p_2}\,\mathrm{d}\lambda \triangleq 1-H^2(P_1, P_2), \tag{10}$$

is called Bhattacharyya coefficient in the field of pattern recognition and affinity in theoretical statistics. We refer to [31] for its formulation in the case of Markov chains and its use for detection.

- *Generalized Matusita distance*

$$r\geqslant 1; f(x)=|1-x^{1/r}|^r, \qquad g(x)=x^{1/r},$$

$$d(P_1, P_2)=\sqrt{\int_{\mathscr{X}} |p_1^{1/r}-p_2^{1/r}|^r\,\mathrm{d}\lambda} \triangleq M_r(P_1, P_2). \tag{11}$$

Notice that, for $r=1$, we get the variational distance and, for $r=2$, the usual Matusita distance, which is equal to $\sqrt{2}H(P_1, P_2)$.

- *Error probability in classification*: It is known that the error probability P_e of the optimal Bayes rule for the classification into 2 classes with a priori probabilities π and $1-\pi$ and where the corresponding densities of the observations are p_1 and p_2, is

$$P_e=\int_{\mathscr{X}} \min[\pi p_1, (1-\pi)p_2]\,\mathrm{d}\lambda. \tag{12}$$

It results that $1-P_e$, which is a way to measure the distance between p_1 and p_2, is of the form (1) with $f(x)=-\min(x, 1-x)$ and $g(x)=x+1$.

Notice that [33] *Patrick and Fisher distance*:

$$d(P_1, P_2)=\sqrt{\int_{\mathscr{X}} (p_1-p_2)^2\,\mathrm{d}\lambda} \tag{13}$$

and *Lissack and Fu distance*

$$(0<\alpha), \qquad d(P_1, P_2)=\int_{\mathscr{X}} |p_1-p_2|^{\alpha}\, d\lambda, \quad (14)$$

are not of the form (1) (except for $\alpha=1$ for the last one).

We will see other examples of spectral distance measures in Section 3.

• *Special case of Gaussian multidimensional laws* $\mathcal{N}(\mu_i, \Sigma_i)$, $(i=1,2)$. This case is investigated in many papers related to the field of pattern recognition. We then get [16, 33]:

(i) Bhattacharyya distance:

$$B(P_1, P_2)=\tfrac{1}{4}(\mu_2-\mu_1)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_2-\mu_1) + \tfrac{1}{2}\operatorname{Log}\frac{|\Sigma_1+\Sigma_2|}{2\sqrt{|\Sigma_1\Sigma_2|}}. \quad (15)$$

(ii) Kullback divergence:

$$J(P_1, P_2)=\tfrac{1}{2}(\mu_2-\mu_1)^T(\Sigma_1^{-1}+\Sigma_2^{-1})(\mu_2-\mu_1) + \tfrac{1}{2}\operatorname{tr}(\Sigma_1^{-1}\Sigma_2+\Sigma_2^{-1}\Sigma_1-2I). \quad (16)$$

When the covariance matrices are identical $\Sigma_1=\Sigma_2=\Sigma$, we get:

(iii) *Mahalanobis distance*:

$$M(P_1, P_2)\triangleq J(P_1, P_2) = 8B(P_1, P_2) = (\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1). \quad (17)$$

2.1.4. Some inequalities

As we said before, the search for bounds of the classification error probability [5, 6, 8, 15], as well as other goals such as feature selection for pattern recognition [8, 33] or signal selection [30, 41], led to various inequalities between P_e and many of the above mentioned distance measures or between the distances themselves. For example [30, 32]:

$$\tfrac{1}{2}[1-\sqrt{1-4\pi(1-\pi)\rho^2}] \leqslant P_e \leqslant \sqrt{\pi(1-\pi)}\,\rho, \quad (18)$$

where ρ is defined in (10);

$$\tfrac{1}{2}\min(\pi, 1-\pi)e^{-J} \leqslant P_e \leqslant \sqrt{\pi(1-\pi)}\left[\frac{J}{4}\right]^{-1/4} \quad (19)$$

and [6]:

$$P_e \leqslant \tfrac{1}{2}-\tfrac{1}{2}V, \quad (20)$$

$$P_e \leqslant \tfrac{1}{2}-\tfrac{1}{2}M_r^r. \quad (21)$$

Other bounds for P_e may be found in [5, 8, 15], and [6] where the case of several classes is also investigated and general bounds are given.

Among the known theoretical inequalities [13], we have

$$H^2(2-H^2)=1-\rho^2, \quad (22)$$

$$e^{-1/2K(P_1, P_2)} \leqslant \rho(P_1, P_2), \quad (23)$$

$$H^2(P_1, P_2) \leqslant V(P_1, P_2) \leqslant H(P_1, P_2)\sqrt{2-H^2(P_1, P_2)}, \quad (24)$$

$$\tfrac{1}{4}e^{K(P_1, P_2)} \leqslant 1-V(P_1, P_2) \leqslant \rho(P_1, P_2). \quad (25)$$

2.2. $\bar{\rho}$ distance

We now introduce another type of distance measure, called the $\bar{\rho}$ or Ornstein distance [20, 21], which is defined both for vector random variables and for processes, and which is theoretically important. An example concerning Gaussian processes will be seen in Section 3.3.5.

In the case of random variables, the definition of this distance is

$$\bar{\rho}(P_1, P_2)=\inf \mathbb{E}|X_1-X_2|^2,$$

where the expectation is with respect to the joint law of the pair (X_1, X_2), and inf is taken over all possible pairs of random variables X_1 and X_2 with probability laws P_1 and P_2 respectively. (Similar definitions do exist for any metric other than the Euclidian one [21].)

When X_i is a scalar zero mean Gaussian variable

with standard deviation σ_i, this reduces to

$$\bar{\rho}(P_1, P_2) = (\sigma_1 - \sigma_2)^2.$$

In the case of general scalar random variables, we also have

$$\bar{\rho}(P_1, P_2) = \int |F_1(x) - F_2(x)| \, dx,$$

where F_1 and F_2 are the distribution functions of the laws P_1 and P_2 respectively. This shows that this distance is not a f-divergence of the form (1).

For random processes, the definition of this distance is quite similar [21]; this distance then measures how two processes look like each other, namely how much one typical realization of one of the processes has to be modified in order to look like to a typical realization of the other one. We shall use this distance in Section 3.3.5.

2.3. Jensen difference and metric on the parameter space

Rao [47] introduces a general approach for associating, to any divergence or dissimilarity measure between parametric probability laws, a metric and appropriate geometries on the parameter space. Before describing this approach, let us present another general divergence measure he introduced earlier under the name of Jensen difference. Using the notations of Section 2.1, let $\mathcal{P}$ be the set of all probability densities on $(\mathcal{X}, \mathcal{F})$. An *entropy* functional is any concave function from $\mathcal{P}$ to $\mathbb{R}$ mapping degenerate densities on zero. Then, for λ_1 and λ_2 positive of sum 1 (usual choice $\frac{1}{2}$ and $\frac{1}{2}$), the *Jensen difference* between two densities p_1 and p_2 is defined to be

$$J(p_1, p_2) = H(\lambda_1 p_1 + \lambda_2 p_2) - \lambda_1 H(p_1) - \lambda_2 H(p_2)$$

and is a measure of dissimilarity (often used in biology). If we consider a family of densities parameterized by $\theta \in \Theta \subset \mathbb{R}^n$, we shall note $J(\theta_1, \theta_2)$. Rao considers the first two terms of the Taylor expansion of $J(\theta, \theta + d\theta)$—which are respectively of order 2 and 3. The first one provides us with an *H-entropy information matrix* (and an H-entropy differential metric) related to the second derivative of J:

$$g_{ij}^H = \left. \frac{\partial^2 J(\theta_1, \theta_2)}{\partial \theta_2^{(i)} \, \partial \theta_2^{(j)}} \right|_{\theta_2 = \theta_1}.$$

If $h(x) = x \operatorname{Log}(x)$, leading to Shannon's entropy, then g_{ij}^h become the elements of the *Fisher information matrix*. A similar general expansion can be derived for the quadratic entropy [47]. If now we consider the Taylor expansion corresponding to the Csiszar f-divergence of Section 2.1, we find that

$$g_{ij}^f(\theta) = f''(1) g_{ij}(\theta),$$

where $g_{ij}(\theta)$ are the elements of the Fisher information matrix. Thus a large class of invariant divergence measures provides the same geometry on the parameter space. However, the third order coefficient may depend on the convex function f, and lead for example to different second order efficiencies of estimates [47].

2.4. General mean distance for classification

For the m-classes classification problem, with a priori probabilities π_i, the error probability $P_e(12)$ becomes

$$P_e = 1 - \int_{\mathcal{X}} \max_i [\pi_i p(x|C_i)] \, dx$$

$$= 1 - \int_{\mathcal{X}} p(x) [\max_i P(C_i|x)] \, dx,$$

where $P(C_i|x)$ is the a posteriori probability of the class C_i given the observation x, and $p(x) = \sum_{i=1}^m \pi_i P(x|C_i)$.

A possible approximation is

$$P_e \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} P_e(C_i, C_j)$$

and for all the pairs (C_i, C_j) the previously mentioned bounds may be used. Another way of getting bounds for P_e was introduced by Van der

Lubbe [6] who defines what he calls the "general mean distance" between the m classes C_i by

$$G_{\alpha,\beta}(C) = \int_{\mathscr{X}} p(x) \left[\sum_{i=1}^{m} P(C_i|x)^{\beta} \right]^{\alpha} dx. \quad (26)$$

This "distance" is symmetric by definition.

This set of distances (26) also contains many known distance measures for pattern recognition, and is related to information measures such as Shannon entropy (also called equivocation) and the quadratic entropy, as can be seen from the following examples.

2.4.1. Examples

The following distance measures were introduced for the derivation of bounds for the error probability P_e which are tightest than Shannon entropy:

$$\sum_i - P(C_i|x) \operatorname{Log} P(C_i|x)$$

- *Devijver Bayesian distance [15]*

(i) $\beta = 2, \alpha = 1$

$$G_{1,2}(C|X) = \int_{\mathscr{X}} p(x) \left[\sum_{i=1}^{m} P(C_i|x)^2 \right] dx \triangleq B(C|X) = 1 - H_2(C|X), \quad (27)$$

where H_2 is the mean conditional quadratic entropy defined from the usual entropy by replacing $-\operatorname{Log} P(C_i|x)$ by $1 - P(C_i|x)$.

(ii) $\beta = 3, \alpha = 1$

$$G_{1,3}(C|X) = \int_{\mathscr{X}} p(x) \left[\sum_{i=1}^{m} P(C_i|x)^3 \right] dx.$$

It can be shown that:

$$G_{1,3}(C|X) = 1 - H_3(C|X), \quad (28)$$

where H_3 is the mean conditional cubic entropy introduced by Chen [8] and defined from the usual entropy by replacing $\operatorname{Log} P(C_i|x)$ by

$$P(C_i|x) - 1 + \tfrac{1}{2}[P(C_i|x) - 1]^2.$$

(iii) $\alpha = 1/\beta, \beta > 1$

$$G_{1/\beta,\beta}(C|X) = \int_{\mathscr{X}} p(x) \left[\sum_{i=1}^{m} P(C_i|x)^{\beta} \right]^{1/\beta} dx$$

is the distance $B'_R(C|X)$ proposed by Trouborst et al. [57]. Many bounds for P_e can be obtained from this class. Among them [6]:

(iv) $\alpha > 0, \beta > 1, 1 \leq \alpha\beta < 1 + \alpha \Rightarrow$

$$1 - G_{\alpha,\beta}(C|X)^{1/\alpha\beta} \leq P_e \leq 1 - G_{\alpha,\beta}(C|X)^{1/\alpha(\beta-1)} \leq 1 - \frac{1}{m^{\alpha}} G_{\alpha,\beta}(C|X).$$

(v) $\alpha > 0, \beta > 1, \alpha\beta \leq 1 \Rightarrow$

$$1 - G_{\alpha,\beta}(C|X) \leq P_e \leq 1 - G_{\alpha,\beta}(C|X)^{1/\alpha(\beta-1)} \leq 1 - \frac{1}{m^{1/\beta}} G_{\alpha,\beta}(C|X)^{1/\alpha\beta}.$$

(vi) $\alpha > 0, \beta > 1, \alpha\beta \geq \alpha + 1 \Rightarrow$

$$1 - G_{\alpha,\beta}(C|X)^{1/\alpha\beta} \leq P_e \leq 1 - G_{\alpha,\beta}(C|X) \leq 1 - \frac{1}{m^{\alpha}} G_{\alpha,\beta}(C|X).$$

(vii) $\lim_{\beta \to \infty} (1 - G_{1/\beta,\beta}(C|X)) = P_e.$

These are generalizations of known bounds. For example, the result corresponding to $\beta = 2$ and $\alpha = 1$ is in [15]. The lower bound corresponding to $\alpha = 1/\beta$, $\beta > 1$ is proved in [57].

2.5. Contrast type distance measures

Another type of distance between laws has been introduced by Poor [45] for robust detection. It is based upon a generalized version of the signal to noise ratio often called contrast.

Given a statistic h for deciding (by comparison to a threshold) between two laws P_1 and P_2, we call the "distance between P_1 and P_2 through the statistics h"

$$S_h(P_1, P_2) = \begin{cases} \dfrac{[\mathbb{E}_2(h) - \mathbb{E}_1(h)]^2}{\operatorname{Var}_1(h)}, & \text{if } \operatorname{Var}_1(h) > 0, \\ 0, & \text{if } \operatorname{Var}_1(h) = 0. \end{cases} \quad (29)$$

If P_1 and P_2 have densities p_1 and p_2, this distance may be written as

$$S_h(P_1, P_2) = \frac{\mathrm{Cov}_1^2(h, \phi)}{\mathrm{Var}_1(h)}, \tag{30}$$

where $\phi = p_2/p_1$. From Schwarz inequality, we have

$$S_h(P_1, P_2) \leq \mathrm{Var}_1(\phi) = S_\phi(P_1, P_2). \tag{31}$$

Notice that S_ϕ belongs to the class (1) with

$$f(x) = (x-1)^2, \qquad g(x) = x. \tag{32}$$

The advantage of this generalized version of the signal-to-noise ratio for robust detection is as follows. The problem of designing robust detectors in terms of risk (in the usual sense of decision theory) amounts to the derivation of a least favorable pair in terms of risk—LMFR; the risk robust detector is then the likelihood ratio of this LMFR pair. The problem is that finding this pair is not always a tractable task. It is thus of interest to search for sub-optimal detectors which are more easily obtainable. It can be shown [45] that, if we define a robustness notion in terms of the distance S, (29), we keep the fact that the robust detector is the likelihood ratio of a least favorable pair in terms of S—LMFS; but we get that such a LMFS pair is often more easily obtainable because it minimises $S_\phi(P_1, P_2)$. Furthermore, this result is also true for the distance S' defined by

$$S'_h(P_1, P_2) = \frac{(\mathbb{E}_2(h))^2}{\mathbb{E}_1(h^2)}, \tag{33}$$

which is also used for detection.

Poor [45] also shows that a LMFR pair is also a pair of closest laws with respect to any f-divergence of the class (1) for any convex continuous f. Furthermore, a LMFR pair is also a LMFS one; but the converse is false.

Finally, referring to Section 2.5 for the local point of view, $\mathrm{Var}_1(\phi) = S_\phi(P_1, P_2)$ plays the same role as Fisher information $I(p) = \int (p')^2/p$ (see Section 2.2) when searching an optimal robust local test for a translation parameter. Indeed, when P_1 has density $p(x)$ and P_2 has density $p(x-\theta)$, where $\theta \to 0$—whence the local terminology—the optimum robust local test is built from the law p which minimizes $I(p)$.

The remainder of this section is devoted to distance measures not between laws but between a law (or a model) and data. This classification is somewhat arbitrary because we could have dealt with this problem above by taking an a priori law as P_1 and an a posteriori or an empirical law as P_2. Nevertheless, we keep this distinction, mainly because of the initial motivations of the hereafter presented tools. Sections 2.6 and 2.7 are even less exhaustive than previous ones. The presented tools will be re-analysed in Section 4 devoted to parametric AR and ARMA models.

2.6. Entropy

In this section, we give the axiomatic derivation of the maximum entropy and minimum cross-entropy (or relative entropy or divergence) principles due to Shore and Johnson [48] because it emphasizes the criteria which lead to these distance measures between models and data already introduced by Kullback [34, 35].

Given a system for which we know:
—an a priori density p;
—constraints I on the "true" unknown density q^* of the form

$$\int q^*(x) a_k(x)\, \mathrm{d}x = 0,$$

or (34)

$$\int q^*(x) c_k(x)\, \mathrm{d}x \geq 0,$$

for known sets of bounded functions a_k and c_k, we investigate the problem of the choice of the best estimate q of q^* knowing the a priori p and the constraints I as given in (34).

We define four axioms which are to be satisfied by the choice criterion and we show that any choice criterion satisfying these axioms is equivalent to the minimization of the relative entropy (or "oriented" divergence or Kullback information as given

in (7)):

$$K(q, p) = \int q(x) \operatorname{Log} \frac{q(x)}{p(x)} \, dx. \tag{35}$$

For this purpose, we introduce the following "information operator" $\circ$: $q = p \circ I$ which associates, to an a priori law p and a set of constraints I on q^*, an a posteriori law q by minimization of a functional H, i.e.,

$$q = p \circ I \Leftrightarrow H(q, p) = \min_{q' \text{ satisfying } I} H(q', p).$$

If there exists another functional H' such that

$$H(q, p) = \min_{q'} H(q', p)$$

$$\Leftrightarrow H'(q, p) = \min_{q'} H'(q', p),$$

H' and H are said to be equivalent, and the operator $\circ$ can be realized using either functional.

The axioms are as follows:

Axiom 1. Unicity. For any p and any I, $q = p \circ I$ is unique;

Axiom 2. Invariance by coordinate transformation. If Γ is a transformation from $\mathscr{X}$ to $\mathscr{Y}$ then:

$$(\Gamma p) \circ (\Gamma I) = \Gamma(p \circ I), \tag{36}$$

where ΓI is the constraint satisfied by the transform of q^*. This means that, if the problem is solved in two different coordinate systems, the two resulting a posteriori densities are related by the coordinate transformation.

Axiom 3. System independence. If $\mathscr{X}_1$ and $\mathscr{X}_2$ are two spaces, with independent a priori densities p_1 and p_2, for which we know the constraints I_1 and I_2, then:

$$(p_1 p_2) \circ (I_1 \wedge I_2) = (p_1 \circ I_1)(p_2 \circ I_2), \tag{37}$$

where $I_1 \wedge I_2$ is the union of the constraints.
This means that the joint a posteriori is the product of the separated a posteriori.

Axiom 4. Subset independence. If $\mathscr{X}$ is a union of disjoint subspaces S_i $(1 \leq i \leq n)$, let $p * S_i$ be the conditional a priori defined by

$$(p * S_i)(x) = \frac{p(x)}{\int_{S_i} p(x') \, dx'}$$

and I_i the constraint on the conditional density $q^* * S_i$, then,

$$(p \circ I) * S_i = (p * S_i) \circ I_i, \tag{38}$$

where $I = I_1 \wedge \cdots \wedge I_n$. (In fact, a stronger condition is imposed [48].)

In order to show that an operator $\circ$ satisfying these four axioms can be realized only by the relative entropy K given in (35), the case of equality constraints in (34) is first investigated (and finally K is shown to work also for inequality constraints). The first step consists in showing that the Axiom 4—see (38) and a special case of Axiom 2—see (36) lead to restricted functionals of the form

$$H(q, p) = \int_{\mathscr{X}} f(q(x), p(x)) \, dx.$$

Then, at the second step, the general case of Axiom 2 is shown to lead to the form

$$H(q, p) = \int_{\mathscr{X}} q(x) f\left(\frac{q(x)}{p(x)}\right) dx. \tag{39}$$

The third step uses Axiom 3 and shows that, if H satisfies the four axioms, H is equivalent to the relative entropy K—see (35). The last step shows that K actually satisfies the four axioms.

This relative entropy minimization principle is successfully used for spectral analysis [49] including in the multidimensional case [29], classification for pattern recognition [50] and many other applications in various domains (see [48]).

2.7. Model validation

We conclude this section with another tool for measuring the distance between a model and data, introduced for signal segmentation and systems monitoring [4]. Let (Y_n) be a controlled Markov process (or more generally a controlled semi-

Markov process) in $\mathbb{R}^k$, the transition probability of which is parameterized by $\theta_* \in \mathbb{R}^d$. Assume that this "true" parameter θ_* is identifiable from the observations Y_n, i.e., there exists a functional H such that the sequence $(\theta_n)_n$ defined by

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n), \tag{40}$$

converges to θ_* (see [4] for precise conditions).

Let θ_0 be a model chosen by the user, and let us investigate the problem of detecting small deviations (local approach as at the end of Section 3.3) $\delta\theta$ with respect to θ_0 using the vector field H as the only statistics. One licit (based upon a central limit theorem) and possible solution consists in considering the random variables

$$Z_k(\theta_0) \triangleq H(\theta_0, Y_k), \tag{41}$$

as if they were independent, *whatever the degree of dependency of the law of the* Y_k is, asymptotically Gaussian distributed, and reflect the small deviation $\delta\theta$ by a change in their mean value. Thus we can use a χ^2 test based upon these Z_k:

$$t = \left(\sum_k Z_k\right)^T R^{-1} \left(\sum_k Z_k\right), \tag{42}$$

where

$$R(\theta_0) = \sum_{n \in Z} \mathrm{Cov}(H(\theta_0, Y_n), H(\theta_0, Y_0))$$

and where the dependency of Z and R in θ_0 has been omitted for simplification.

In (42), we assume that, if

$$h(\theta) \triangleq \mathbb{E}_\theta(H(\theta, Y_n)),$$

then $\dot{h}(\theta)$ (i.e., the derivative of h) is invertible. If this is not the case, see [4].

Equation (42) is clearly a way for measuring the agreement (or the deviation) between the model θ_0 and the observations (Y_n). This way is obviously not the unique possible one. Another one, more classical, consists in running the algorithm (40) and using a χ^2 test of the form

$$(\theta_n - \theta_0)^T \Sigma^{-1} (\theta_n - \theta_0) \gtrless \lambda, \tag{43}$$

using the fact that $\theta_n - \theta_0$ is asymptotically Gaussian distributed with zero mean. But it turns out that $\theta_n - \theta_0$ has a quite complex dynamics (Gaussian-Markov process of first order), and its temporal dependency structure, which is not taken into account in (43), is better reflected in (42) which is probably more efficient.

Finally, we refer to [52] for a special use of Kullback information—see (7)—between conditional laws for multivariable input/output model validation.

3. Spectral distance measures

In this section, we are interested in spectral distance measures, namely in distances between processes based upon their second order properties. Some of these distance measures have already been introduced in the previous section, but by far not all of them, and thus it is obviously interesting to present together all the possible distances. We recall that the formulation of these distances, when the spectra are represented by parametric AR or ARMA models, will be addressed in the next section, together with the questions related to parameter estimation.

The key references for this problem are without doubt to be found in [18-20, 32, 40, 44].

3.1. Preliminary remarks

Following [19], we shall use the following notations. Let $s(\lambda)$ be a (energy or power) spectral density corresponding to a scalar signal. λ varies from $-\pi$ to π, where we assume that π is half of the sampling frequency of the signal. s is a positive even function, the Fourier coefficients of which define an autocorrelation sequence:

$$\begin{aligned} s(\lambda) &= \sum_{n \in Z} r(n)\, \mathrm{e}^{-jn\lambda}, \\ r(n) &= \int_{-\pi}^{\pi} \frac{1}{2\pi} s(\lambda)\, \mathrm{e}^{jn\lambda}\, \mathrm{d}\lambda. \end{aligned} \tag{44}$$

Following the terminology introduced in the introduction, if r is the theoretical autocorrelation function, the following distance measures will be distances between laws of processes, and if r is

the empirical autocorrelation, the distances will be distances between signals.

Let $R_N(s)$ be the Toeplitz $(N+1)\times(N+1)$ matrix, the (k, j)th element of which is $r(k-j)$, $0 \leq k, j \leq N$. We shall use several fundamental properties of R_N [19, 22]. $|R_N|$ will denote the determinant of R_N.

For each p, a Toeplitz form can be associated to the spectral density s by

$$T_p(a) \triangleq \int_{-\pi}^{\pi} \frac{1}{2\pi} \left| \sum_{k=0}^{p} a_k \, e^{jk\lambda} \right|^2 s(\lambda) \, d\lambda$$

$$= \sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l r(k-l)$$

$$= a^T R_p(s) a \qquad (45a)$$

where $a^T = (a_0, a_1, \ldots, a_p)$ is real.

A numerically convenient form is

$$T_p(a) = r(0) r_a(0) + 2 \sum_{k=1}^{p} r(k) r_a(k), \qquad (45b)$$

where

$$r_a(k) \triangleq \sum_{l=0}^{p-k} a_l a_{l+k}, \quad 0 \leq k \leq p.$$

We shall see later that this Toeplitz form directly appears in spectral distance measures, and especially in distances between a model a and a signal summarized in its covariances R_N.

Let

$$A(z) = \sum_{k=0}^{p} a_k z^{-k}, \qquad (46)$$

then

$$T_p(a) = \int_{-\pi}^{\pi} \frac{1}{2\pi} |A(e^{j\lambda})|^2 s(\lambda) \, d\lambda.$$

And letting

$$\sigma_s^2(p) = \min_{\substack{a \\ (a_0=1)}} T_p(a),$$

then [22]

$$\sigma_s^2(p) = \frac{|R_p(s)|}{|R_{p-1}(s)|}$$

and the minimizing polynomial $A(z)$ may be analytically expressed in terms of orthogonal polynomials (cf. Levinson algorithm). Let $A_p(z)$ be the pth order polynomial with $a_0 = 1$ which minimizes (45). This polynomial $A_p(z)$ together with $\sigma_s^2(p)$ may be used to model the spectral density $s(\lambda)$. Actually, for any polynomial

$$G(z) = \sum_{k=0}^{n} g_k z^{-k},$$

and we can write

$$T_p(g) \triangleq \int_{-\pi}^{\pi} \frac{1}{2\pi} |G(e^{j\lambda})|^2 s(\lambda) \, d\lambda$$

$$= \int_{-\pi}^{\pi} \frac{1}{2\pi} |G(e^{j\lambda})|^2 \frac{\sigma_s^2(p)}{|A_p(e^{j\lambda})|^2} \, d\lambda.$$

Furthermore, let [22]

$$\sigma_s^2 \triangleq \lim_{p \to \infty} \sigma_s^2(p)$$

$$= \exp\left[\int_{-\pi}^{\pi} \frac{1}{2\pi} \operatorname{Log}(s(\lambda)) \, d\lambda \right] \qquad (47)$$

and let us consider the following spectral factorization:

$$\frac{1}{s(\lambda)} = \frac{|A(e^{j\lambda})|^2}{\sigma_s^2}, \qquad (48)$$

where $A(z) = \lim_p A_p(z)$ has no zero on or outside the unit circle. We shall call σ_s^2/A the (infinite) autoregressive model of s, and $1/A$ the normalized AR model.

Most of the spectral distance measures which we shall consider will be in terms of L_q norms, i.e.,

$$\|s\|_q = \left[\int_{-\pi}^{\pi} \frac{1}{2\pi} |s(\lambda)|^q \, d\lambda \right]^{1/q}, \qquad (49)$$

which satisfies

$$\|s\|_{q_1} \leq \|s\|_{q_2} \quad \text{for} \quad 0 < q_1 \leq q_2.$$

If s is continuous, $\|s\|_\infty$ exists and is the maximum magnitude of s.

3.2. Spectral distance measures and equivalences

Spectral distances between two spectral densities s_1 and s_2 may be measured with the aid

of L_q norms of their difference, i.e.,

$$d(s_1, s_2) = \|s_1 - s_2\|_q.$$

These distances are "true" distances in the sense that they satisfy the symmetry property and the triangular inequality. We saw examples of such distance measures in Section 2.1. However, the spectral distances which will be used here are functions of the *difference between the log-spectra*, i.e., of the ratio between the spectra:

$$d(s_1, s_2) = d\left(1, \frac{s_2}{s_1}\right) = d\left(\frac{s_1}{s_2}, 1\right) \tag{50}$$

for obvious requirements of invariance with respect to the measurement scale.

For a given distance d, we shall use two types of scaling [19]. A *gain normalized distance measure* is defined by

$$d^*(s_1, s_2) = d\left[\frac{s_1}{\sigma_1^2}, \frac{s_2}{\sigma_2^2}\right], \tag{51}$$

where σ_1 and σ_2 are defined in (47) and correspond to s_1 and s_2 respectively. This distance is useful for separating the effects of the normalized models and the gains.

A *gain optimized distance measure* is defined by

$$d'(s_1, s_2) \triangleq \min_{\alpha \geq 0} d(s_1, \alpha s_2). \tag{52}$$

By definition, $d(s_1, s_2) \geq d'(s_1, s_2)$.

Notice that the usual spectral distance measures are easily defined in the spectral domain, but are most of the time numerically computed without reference to this domain.

As there exist many spectral distance measures d (and d' and d^* defined above are ways to introduce variants!), it is important to know when they are equivalent. Intuitively, two distances are equivalent if the results obtained for a given application with either of them are qualitatively the same. More precisely, following again [19], we define two types of equivalence. The first one is the usual equivalence for metrics. The second one is a convenient equivalence for coding and classification problems (search of nearest neighbor).

A distance d_1 is said to be *stronger* than a distance d_2, and we write

$$d_1 \Rightarrow d_2,$$

if a small distance d_1 implies a small distance d_2. d_1 and d_2 are said to be *equivalent* if each is stronger than the other.

Let us now consider the problem of finding a nearest neighbor (NN), i.e., of a representation $\hat{s}$ of s in a particular set which minimizes a distance. d_1 and d_2 are *NN-equivalent* if the two corresponding functions $s \mapsto \hat{s}$ are identical, whatever the representation set is. This equivalence can be very useful in practice because it allows to use the simplest NN-equivalent distance for the computations.

If two distances d_1 and d_2 are equivalent in both senses, they are said to be *completely equivalent* and we write

$$d_1 \langle \equiv \rangle d_2.$$

From (51), (50), (52), we get

$$d^*(s_1, s_2) \geq d'(s_1, s_2).$$

Thus *d and d^* are stronger than d'.*

3.3. Main spectral distance measures

3.3.1. Log spectral deviation

This measure is probably the oldest one in speech processing, and is defined by the L_q norm of the difference of the logarithms of the spectra:

$$d_q(s_1, s_2) = \|\operatorname{Log} s_1 - \operatorname{Log} s_2\|_q = \left\|\operatorname{Log} \frac{s_1}{s_2}\right\|_q. \tag{53}$$

The more common choices are:

- $q = 1$: mean absolute distance,
- $q = 2$: mean quadratic distance (rms),
- $q = \infty$: maximum deviation.

We have $d_\infty \geq d_2 \geq d_1$. These distances satisfy the symmetry property and the triangular inequality. They are directly related to decibel variations in

the log spectral domain by the factor $10/\text{Log } 10 = 4.34$. The L_2 norm is the most popular because the most easily computable. Approximations will be mentioned in the next section. Moreover, it turns out to be experimentally close to L_∞ [18], at least when the spectra are estimated via Fourier transform.

3.3.2. Itakura-Saïto distance [27]

The Itakura-Saïto distance is defined by

$$d_{IS}(s_1, s_2) = \left\| \frac{s_1}{s_2} - \text{Log}\frac{s_1}{s_2} - 1 \right\|_1 \tag{54}$$

and is also called "error matching measure".

As $u - \text{Log } u - 1 \geq 0$, we also have

$$d_{IS}(s_1, s_2) = \int_{-\pi}^{\pi} \frac{1}{2\pi}\frac{s_1}{s_2}\,d\lambda - \text{Log}\frac{\sigma_1^2}{\sigma_2^2} - 1. \tag{55}$$

Using the expansion of $u = \exp(\text{Log } u)$, it can be shown that d_{IS} is an approximation for $\frac{1}{2}d_2^2$ for "small" distances.

On the other hand, by Jensen inequality, we have

$$d_{IS}(s_1, s_2) \geq d_{IS}(\sigma_1^2, \sigma_2^2).$$

For s_2 of the form

$$s_2(\lambda) = \frac{\sigma_2^2}{|A_2(e^{j\lambda})|^2},$$

where A_2 is causal of order p—specifically if we want to solve the problem of linear prediction of s_1—from (55) and (45) we conclude that:

$$d_{IS}(s_1, s_2) = \frac{1}{\sigma_2^2} T_p^{(1)}(a_2) - \text{Log}\frac{\sigma_1^2}{\sigma_2^2} - 1. \tag{56}$$

We shall come back to this expression in the next section.

Another form of the Itakura-Saïto distance has actually already been mentioned in the last section. Consider the Kullback information (7) for Gaussian processes [44]:

$$K_N(s_1, s_2) = \tfrac{1}{2}\,\text{Log}\frac{|R_N(s_1)|}{|R_N(s_2)|} + \tfrac{1}{2}\,\text{tr}[R_N(s_1)R_N^{-1}(s_2)] - \tfrac{1}{2}N. \tag{57}$$

It can be shown that [44]

$$K(s_1, s_2) \triangleq \lim_N \frac{1}{N} K_N(s_1, s_2) = \tfrac{1}{2}\, d_{IS}(s_1, s_2). \tag{58}$$

In other words, Itakura-Saïto distance is equal to two times the asymptotic Kullback information under Gaussian hypothesis. This technique has been successfully tested for classifying non-Gaussian data for the purpose of recognition of EEG signals [17]. Furthermore, d_{IS}, even though non symmetrical, is well suited to quantification, classification, recognition, and detection problems, at least in the domain of speech processing [19]. This is also the case for classification [25] and recognition of EEG signals once more, for which in [24] Kullback distance, Kullback divergence and Bhattacharrya distance have been compared.

3.3.3. Itakura distance

$$d_I(s_1, s_2) \triangleq d'_{IS}(s_1, s_2). \tag{59}$$

From (55), we get

$$d_I(s_1, s_2) = \text{Log} \int_{-\pi}^{\pi} \frac{1}{2\pi}\frac{s_1/\sigma_1^2}{s_2/\sigma_2^2}\,d\lambda \tag{60}$$

and

$$d_{IS}(s_1, s_2) = \frac{\sigma_1^2}{\sigma_2^2}\exp[d_I(s_1, s_2)] - \text{Log}\frac{\sigma_1^2}{\sigma_2^2} - 1. \tag{61}$$

Using (48) as model for s_1 and s_2, we get

$$d_I(s_1, s_2) = \text{Log} \int_{-\pi}^{\pi} \frac{1}{2\pi}\left|\frac{A_2}{A_1}\right|^2 d\lambda = \text{Log}\left[\left\|\frac{A_2}{A_1}\right\|_2^2\right] \tag{62}$$

This distance is also called log likelihood ratio because of its asymptotic expression in the Gaussian case [26, 55]. We refer to the next section for additional details.

3.3.4. Model distance measure

The model distance measure also introduced by Itakura [26] is defined by

$$d_{\mathrm{m}}^{*}(s_1, s_2) \triangleq \left\| 1 - \frac{A_2}{A_1} \right\|_2^2 \tag{63}$$

where A_1 and A_2 are the normalized AR models for s_1 and s_2. It can be shown that [19]:

$$\begin{aligned} d_{\mathrm{m}}^{*}(s_1, s_2) &= \left\| \frac{A_2}{A_1} \right\|_2^2 - 1 \\ &= \exp(d_{\mathrm{I}}(s_1, s_2)) - 1 \end{aligned} \tag{64}$$

and thus d_{m}^{*} and d_{I} are completely equivalent. We also have

$$d_{\mathrm{m}}^{*} = d_{\mathrm{IS}}^{*}.$$

This distance was introduced as an approximation of d_{I} for d_{I} small (see (64)). It is always an upper bound for d_{I}. It is called a model distance measure because it measures how the normalized models or filters A_1 and A_2 are close to being inverse of each other (see next section).

A similar unnormalized model distance is given by

$$\begin{aligned} d_{\mathrm{m}}(s_1, s_2) &= \left\| 1 - \frac{\sigma_1/A_1}{\sigma_2/A_2} \right\|_2^2 \\ &= \frac{\sigma_1^2}{\sigma_2^2} d_{\mathrm{m}}^{*}(s_1, s_2) + \left(1 - \frac{\sigma_1^2}{\sigma_2^2} \right)^2 \\ &= \frac{\sigma_1^2}{\sigma_2^2} d_{\mathrm{m}}^{*}(s_1, s_2) + d_{\mathrm{m}}(\sigma_1^2, \sigma_2^2) \end{aligned} \tag{65}$$

But

$$d_{\mathrm{IS}}(s_1, s_2) = \frac{\sigma_1^2}{\sigma_2^2} d_{\mathrm{m}}^{*}(s_1, s_2) + d_{\mathrm{IS}}(\sigma_1^2, \sigma_2^2),$$

thus

$$d_{\mathrm{IS}} \Leftrightarrow d_{\mathrm{m}}.$$

However d_{IS} and d_{m} are not NN-equivalent. The optimization of d_{m} according to (52) gives

$$d'_{\mathrm{m}}(s_1, s_2) = 1 - 1 \Big/ \left\| \frac{A_1}{A_2} \right\|_2^2, \tag{66}$$

which can be shown to be a monotonic function of d_{m}^{*}, and thus

$$d'_{\mathrm{m}} \langle \equiv \rangle d_{\mathrm{m}}^{*}.$$

3.3.5. Symmetrized distance measures

A spectral distance measure d can be symmetrized by defining for $q \geq 1$:

$$d^{(q)}(s_1, s_2) = \tfrac{1}{2}(d(s_1, s_2)^q + d(s_2, s_1)^q)^{1/q}. \tag{67}$$

Note that $d^{(q)}$ is stronger than d.

A symmetrized version of the Itakura-Saito distance was introduced in [18] and defined by

$$d_{\cosh}(s_1, s_2) \triangleq d_{\mathrm{IS}}^{(1)}(s_1, s_2), \tag{68}$$

where the terminology cosh (of the spectral difference measured on a logarithmic scale) comes from (54). $d_{\cosh}$ is related to a decibel scale [18] with the aid of the quantity D such that $\cosh(D) - 1 = d_{\cosh}$, i.e.,

$$D = \mathrm{Log}(1 + d_{\cosh} + \sqrt{d_{\cosh}(2 + d_{\cosh})}).$$

From (58), we conclude that Kullback divergence:

$$J_N(s_1, s_2) \triangleq K_N(s_1, s_2) + K_N(s_2, s_1)$$

satisfies:

$$\lim_{N \to \infty} \frac{1}{N} J_N(s_1, s_2) = d_{\cosh}(s_1, s_2). \tag{69}$$

It can also be shown that [20, 40]:

$$2 d_{\cosh}(s_1, s_2) = \bar{\rho}(Y^{(1)}, Y^{(2)}),$$

where $Y^{(1)}$ and $Y^{(2)}$ are two Gaussian processes with spectral densities s_1/s_2 and s_2/s_1 respectively, and where $\bar{\rho}$ is the Ornstein distance [20, 21] between processes already mentioned in Section 2.2. For Gaussian processes $X^{(1)}$ and $X^{(2)}$ with spectral densities s_1 and s_2, we have [20, 40]:

$$\bar{\rho}(X^{(1)}, X^{(2)}) = \| \sqrt{s_1} - \sqrt{s_2} \|_2^2.$$

This leads to the following (simple) relationship:

$$2 d_{\cosh}(s_1, s_2) = \left\| \sqrt{\frac{s_1}{s_2}} - \sqrt{\frac{s_2}{s_1}} \right\|_2^2. \tag{70}$$

3.3.6. Summary of the equivalences

Many other symmetric distance measures may be defined by symmetrizing the previously mentioned distances or by gain optimizing or gain normalizing the above mentioned symmetric distances. Recall that we always have

$$d^{(1)} \Rightarrow d, \qquad d^{*} \Rightarrow d', \qquad d \Rightarrow d'.$$

The known equivalences between the above distance measures are summarized in the diagram of Table 1 [19]. Other results are described in [40] together with their consequences on robustness issues of linear predictive coding.

Table 1
Summary of the equivalences [19]

$$\begin{array}{ccccccc} d_{\mathrm{I}}^{(1)} & \Leftrightarrow & d'_{\cosh} & \Leftrightarrow & d_{\mathrm{m}}^{(1)*} & \Rightarrow & d_{\mathrm{I}} = d'_{\mathrm{IS}} \Leftrightarrow d_{\mathrm{m}}^{*} = d_{\mathrm{IS}}^{*} \Leftrightarrow d'_{\mathrm{m}} \\ & & \Uparrow & & & & \Uparrow \\ d_{\cosh} = d_{\mathrm{IS}}^{(1)} & \Leftrightarrow & d_{\mathrm{m}}^{(1)} & \Rightarrow & d_{\mathrm{IS}} & \Leftrightarrow & d_{\mathrm{m}} \\ \Downarrow & & & & & & \\ d_2 & & & & & & \end{array}$$

3.3.7. The case of Gaussian processes

In [32] closed form numerically computable formulas were obtained for Bhattacharrya distance, Chernoff distance, Kullback distance and Kullback divergence between two r-dimensional Gaussian processes $Y^{(1)}$ and $Y^{(2)}$. These expressions are in terms of the two spectral densities matrices $S_1(\lambda)$ and $S_2(\lambda)$ corresponding to the two covariance matrices sequences, and of the spectral density matrix $M(\lambda)$ of the difference between the process means. For example [32]:

$$\begin{aligned} 2K(Y^{(2)}, Y^{(1)}) = \int_{-\pi}^{\pi} \frac{1}{2\pi} (\operatorname{tr} S_1^{-1}(\lambda) \\ \times [S_2(\lambda) - S_1(\lambda)] \\ - \log S_1^{-1}(\lambda) S_2(\lambda)) \mathrm{d}\lambda \\ + \int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{k=1}^{r} \sum_{j=1}^{r} \\ \times m_{kj}(\lambda) s_{kj}(0, \lambda)\, \mathrm{d}\lambda, \end{aligned}$$

where

$$M(\lambda) = (m_{kj}(\lambda))_{1 \leq k, j \leq r},$$

$$[(1-t)S_1(\lambda) + tS_2(\lambda)]^{-1} = (s_{kj}(t, \lambda))_{1 \leq k, j \leq r}.$$

In the scalar case, other spectral expressions of information measures may be found in [51] together with their relation to order criteria.

4. Parametric spectral distance measures

In this section, we investigate the practically important special case where the spectra are described by AR or ARMA parametric models. We describe the useful expressions for many previously mentioned distance measures. The relationships between some of them together with the possible problems related to the interaction between these distances and the choice of parameters (and the way by which they have been estimated) are also addressed. We mention some variants still currently introduced for speech recognition systems performance improvement. Finally, we present some qualitative results from comparative studies for distance measures.

4.1. L_2 norm and cepstral distance

In Section 3.3.1, we indicated that the L_2 norm of the log-spectra difference is a commonly used distance measure especially for speech processing. However the main drawback of this distance d_2 is of computational nature, because it requires two FFT, two logarithms and one summation. In this section, we show how it can be efficiently approximated by an Euclidian distance; the cepstral distance.

Given a pth order minimum phase filter, namely:

$$A(z) = \sum_{k=0}^{p} a_k z^{-k}, \tag{71}$$

with $a_0 = 1$, having all its the roots inside the unit circle, we define the *cepstral coefficients* [18] to be the coefficients of the Taylor expansion of the

logarithm of the filter transfer function, i.e.

$$\text{Log } A(z) = -\sum_{k=1}^{\infty} c_k z^{-k}. \tag{72}$$

They are also the Fourier coefficients of the log-spectrum, because

$$\text{Log } \frac{\sigma^2}{|A(e^{j\lambda})|^2} = \sum_{k=-\infty}^{\infty} c_k e^{-jk\lambda}, \tag{73}$$

where

$$c_0 = \text{Log } \sigma^2 \quad \text{and} \quad c_{-k} = c_k. \tag{74}$$

These cepstral coefficients may be estimated in two ways. The traditional first one consists in two FFT starting from the filter impulse response

$$H(z) \triangleq \frac{1}{A(z)} = \sum_{n=0}^{\infty} h_n z^{-n}. \tag{75}$$

For a transfer function with poles only, the cepstrum can be obtained directly from the impulse response coefficients h_n by

$$\begin{aligned} c_n &= \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) h_k c_{n-k} + h_n, \quad n > 1, \\ c_1 &= h_1, \end{aligned} \tag{76}$$

or from the linear prediction coefficients by

$$\begin{aligned} c_1 &= -a_1 \\ c_n &= -a_n - \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, \quad 1 < n \leq p, \\ c_n &= -\sum_{k=1}^{p} \frac{n-k}{n} c_{n-k} a_k, \quad p+1 \leq n. \end{aligned} \tag{77}$$

Notice that the two resulting cepstra ("Fourier" and "parametric") are not identical [2]. More generally, the cepstrum of a minimum phase rational transfer function can be defined [43]. In this case, the cepstral coefficients can be expressed as a function of the poles $(z_k)_{1 \leq k \leq p}$ and the zeroes $(w_k)_{1 \leq k \leq q}$ [38, 43]:

$$c_n = -\frac{1}{n}\left[\sum_{k=1}^{p} z_k^n - \sum_{j=1}^{q} w_j^n\right], \quad n > 0. \tag{78}$$

Using (73) and applying Parseval formula to d_2 (53), we get

$$\begin{aligned} d_2^2 &= \sum_{k=-\infty}^{+\infty} (c_k^{(1)} - c_k^{(2)})^2 \\ &= (c_0^{(1)} - c_0^{(2)})^2 + 2\sum_{k=1}^{\infty} (c_k^{(1)} - c_k^{(2)})^2, \end{aligned} \tag{79}$$

where the $c_k^{(i)}$ $(i = 1, 2)$ are the cepstral coefficients associated to the spectral density s_i. Furthermore, the finite sums:

$$d^2(L) = \sum_{k=-L}^{L} (c_k^{(1)} - c_k^{(2)})^2, \quad L \geq p \tag{80}$$

can be shown to be positive definite and to converge, when $L \to \infty$, towards d_2^2. Moreover, experiments with speech signals [18] have shown that, for small values of L, d(L) is closed to d_2. The usual values for L are p and $2p$.

Therefore, as far as spectral distance measures are concerned, the "good" Euclidian distance is between the cepstral coefficients c_n (77) (and not between the autoregressive coefficients a_n!). Furthermore, the cepstral distance is experimentally better than the Euclidian distance between the reflection coefficients [2, 58]. We shall discuss this point further in Section 4.4. Moreover, from (78) and (80), we conclude that, for causal spectra

$$d^2(L) = \sum_{k=-L}^{L} \frac{1}{k^2}\left[\sum_{i=1}^{p} (z_i^{(2)k} - z_i^{(1)k})\right]^2, \tag{81}$$

where $z_i^{(j)}$ $(j = 1, 2; \quad i = 1, p)$ are the poles of the spectrum s_j. This shows that one can be vary far from a true spectral distance when one tries—in an intuitively "natural" way—to measure the deviation between two spectra with the aid of an Euclidian or absolute value distance between the poles (or the Fourier spectrum lines).

A last important remark about the cepstral distance concerns the interaction between parameter estimation and distance. Actually, it seems that the distance $d(L)$ (80) is not to be used when the AR

coefficients (a_k) used in (77) are estimated with the autocorrelation method [3].

4.2. Distances d_{IS} and d_I

We now consider the parametric formulation of the Itakura-Saïto distance d_{IS} as given in (54) and Itakura distance d_I as given in (60). We also present a variant of d_I.

A parametric expression of the distance d_{IS} has already been given in (56). From (61) and (56) we get for d_I:

$$d_I(s_1, s_2) = \text{Log}\frac{T_p^{(1)}(a_2)}{\sigma_1^2} = \text{Log}\frac{a_2^T R_p^{(1)} a_2}{\sigma_1^2}. \quad (82)$$

Notice that, if from (62) d_I is a distance between models, from (82) it is rather a distance between a model a_2 and a signal y summarized in its autocorrelation matrix $R_p^{(1)}$ and "residual energy" σ_1^2 (47).

The dissymmetry of $T_p^{(i)}(a_j)$ with respect to i and j—embarrassing for solving the problem inverse of linear prediction—is also met in d_I and d_{IS} and reflects nothing but the known dissymmetry of Kullback distance (see (58)).

The distance d_I is widely used in speech recognition systems, but its main drawback—as for many other distance measures—is its lack of robustness in presence of noise, especially if the learning step has been done with non noisy speech signals. For this reason, a weighted Itakura distance was recently introduced [53]. The weighting is done with the aid of Atal perceptual filter:

$$d_{wI} = \text{Log}\int_{-\pi}^{\pi}\frac{1}{2\pi}\frac{1}{|A_1'(e^{j\lambda})|^2}\frac{|A_2(e^{j\lambda})|^2}{|A_1(e^{j\lambda})|^2}\,d\lambda \quad (83)$$

where

$$A_1'(e^{j\lambda}) = A_1(\alpha e^{j\lambda}) = 1 + \alpha a_1^{(1)} e^{-j\lambda} + \cdots + \alpha^p a_p^{(1)} e^{-jp\lambda}$$

and $0 \leq \alpha \leq 1$ allows to increase the band pass width. This filter has also been used in order to improve the performances of the cepstral distance (80) [39].

4.3. Some other distance measures

4.3.1. Variants of the cepstral distance

[28, 39, 59] introduced several distance measures based upon the derivative of the phase spectrum—namely the group delay—rather than the logarithm of the spectrum. Let us consider the Taylor expansion of the phase $\phi(z)$ of Log $A(z)$. From (72)

$$\phi(e^{-j\lambda}) = \sum_{k=1}^{\infty} c_k \sin(k\lambda). \quad (84)$$

Thus the expansion of the group delay is

$$\phi'(\lambda) = \frac{d\phi(e^{-j\lambda})}{d\lambda} = \sum_{k=1}^{\infty} k c_k \cos(k\lambda). \quad (85)$$

Introducing as a new spectral distance the quantity

$$\tilde{d} = \|\phi_1' - \phi_2'\|_2,$$

Yegnanarayana [59] suggests to use the following Euclidian distance:

$$\tilde{d}^2(L) = \sum_{k=1}^{L} k^2 (c_k^{(1)} - c_k^{(2)})^2, \quad (86)$$

where L has to be chosen higher than the order p because the convergence of the series (86) is slower than that of (80).

More generally, in [28] Itakura recently suggested to use an Euclidian distance based upon a "smoothed" group delay, namely upon $w_k c_k$, where

$$w_k = k^s e^{-k^2/2\tau^2}, \quad s \geq 0.$$

Finally and even more recently [39] more robust spectral distances based upon the cosine of the angle between two cepstral coefficients arrays, with c_0 excluded, have been introduced:

$$|C_1|^2(1 - \cos^2\beta)$$

and

$$|C_1|^\alpha(1 - \cos\beta), \quad \alpha = 0, 1, 2$$

where

$$\cos\beta = \frac{C_1^T C_2}{|C_1||C_2|},$$

where C_j is the vector of the cepstral coefficients $c_k^{(j)}$ $(k>0, j=1,2)$. Other weightings have also been investigated in [9].

4.3.2. A divergence between conditional laws

Finally, let us mention a special distance used for signal segmentation [3]. This "distance" is based upon Kullback divergence between the *conditional laws* $p_j(y_n | y_{n-1}, \ldots, y_{n-p})$ of the observed signal (y_n) computed for two estimated Gaussian $AR(p)$ models $(j=1,2)$ long-term and short-term respectively. The reasons for this choice are explained in [3]. The particular point here is that the resulting distance is

$$-\frac{e_n^{(1)}e_n^{(2)}}{\sigma_2^2}+\frac{1}{2}\left[1+\frac{\sigma_1^2}{\sigma_2^2}\right]\frac{e_n^{(1)^2}}{\sigma_1^2}-\frac{1}{2}+\frac{\sigma_1^2}{2\sigma_2^2} \qquad (87)$$

where the $e_n^{(j)}$ are the innovations of the two filters. Thus this "distance" is actually a random variable, which turns out to have high sensitivity with respect to spectral changes in speech signals.

An interesting practical property of (87) is that the quality of the resulting segmentation is better when the identification methods for computing $e_n^{(j)}$ and σ_j^2 are approximated least squares than when they are exact. We have no theoretical explanation for this strange parameters/distance interaction (see also the remark at the end of Section 4.1).

4.4. Comparisons of distances and parametrizations

Many comparative studies for distances, and also for choices of parametric representations, have been conducted in the field of speech recognition, but also in other domains [24, 36, 58]. The oldest ones are probably due to Atal [2] who already noticed that the cepstral distance (80) is better than the Euclidian distance between the reflection coefficients. These results were corroborated for example in [23], and [14] where the cepstrum (73) obtained by Fourier analysis—in the so-called mel scale—seemed to lead to a better distance measure than the parametric cepstrum computed by (77), maybe because of consonants; moreover, in this study, the cepstral distance d_2 appeared to be better than the Itakura distance d_1. (Notice that they cannot be compared in the table of Section 3.3.6). Similar conclusions concerning the cepstrum have been obtained in recent work [9]. On the other hand, in [36] several distances are compared for varying AR orders in an underwater acoustics application. The best selected distance is $d_1^{(1)}$.

Other comparisons have been done [42], with different weighting variants introduced for speech signals ("spectral slope", ...).

Recall that the variants of d_1 or d_2 introduced respectively in [53] and [28, 39] have been compared to the original distance d_1 or d_2.

It is not easy to draw a synthetic picture from these comparative studies, even for the only domain of speech recognition, because their experimental conditions are highly variable (sets of reference signals used for learning and of test signals used for recognition).

Recall that the most fundamental comparative analysis has been conducted by Matsuyama [19, 40] and is partly summarized in Table 1.

5. Conclusion

It is quite difficult to give strong and definite recommendations about the choice of a distance measure for a given particular application. The fact that, even in the field of speech processing where this question has been investigated probably for the longest time, new variants of spectral distance measures still emerge each year, shows that this problem of choice is quite complex. Any quantitative result about frequencies of use of given distance measures in the literature would probably be too much biased by the selection of the papers itself. However, we think that from the above whole set of studies concerning distance measures arise some elements leading to a kind of conclusion about the distance measures to be preferred in practice. Actually, Table 1 shows that the strongest distance is $d_{\cosh}$, which is a Kullback divergence from (69), and also a $\bar{\rho}$ distance. On

the other hand, it is known that deep theoretical results in Ergodic Theory and in Statistics have been shown using the $\bar{\rho}$ and the Hellinger distances. Finally, in the pattern recognition literature (see references below), the preferred distances are Mahalanobis, Matusita and Bhattacharyya distances, which are, from (17), (11) for $r=2$, and (10), related to the Hellinger distance (and also Kullback divergence). Thus, we may conclude that Kullback divergence J (8) and Hellinger distance H^2 (6), which take a key part for proving complex theoretical results as well as solving applied problems, have to be preferred in practice. Furthermore, we think it quite stimulating to find out that the same tools are preferred by theoreticians and practitioners. In addition to J and H, we also recommend to use d_2 (79) in practice, because of its Euclidian nature and thus its computational simplicity. Recall that, from this computational point of view, the NN-equivalence (see Section 3.2) between distance measures is of key interest.

Finally, let us outline that the question of the order of magnitude of the distances to be measured in a particular application may be a key issue in the selection of a distance measure. Recall that, according to [47] for example (see Section 2.2), many f-divergence measures (1) "locally"—namely for small distances—provide the same "information" on the parameter space, but this is no longer true if "large" distances are to be measured. This nonuniform efficiency of distance measures is a well known issue in speech processing again.

Acknowledgement

The author wishes to thank Albert Benveniste, Jean Deshayes and the two anonymous referees for many comments which helped improving earlier versions of the paper.

References

[1] S.M. Ali and D. Silvey, "A general class of coefficients of divergence of one distribution from another", *J. Roy. Stat. Soc. B*, Vol. 28, No. 1, 1966, pp. 131-142.

[2] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Am.*, Vol. 55, No. 6, June 1974, pp. 1304-1312.

[3] M. Basseville, "The two-models approach for the on-line detection of changes in *AR* processes", in: M. Basseville, A. Benveniste, eds., *Detection of Abrupt Changes in Signals and Dynamical Systems*, LNCIS No. 77, Springer-Verlag, 1986.

[4] A. Benveniste, M. Basseville and G. Moustakides, "The asymptotic local approach to change detection and model validation", *IEEE Trans. Autom. Control*, Vol. AC-32, No. 7, July 1987, pp. 583-592.

[5] D.E. Boekee and J.C. Ruitenbeek, "A class of lower bounds on the Bayesian probability of error", *Inf. Sci.*, Vol. 25, 1981, pp. 21-25.

[6] D.E. Boekee and J.C.A. Van Der Lubbe, "Some aspects of error bounds in feature selection", *Pattern Recognition*, Vol. 11, 1979, pp. 353-360.

[7] J. Burbea and C.R. Rao, "On the convexity of some divergence measures based on entropy functions", *IEEE Trans. Inf. Theory*, Vol. IT-28, No. 3, May 1982, pp. 489-495.

[8] C.H. Chen, "On information and distance measures, error bounds, and feature selection", *Inf. Sci.*, Vol. 10, 1976, pp. 159-173.

[9] J.P. Cordeau, Un système de reconnaissance—Analyse de quelques métriques, Stage Report, ENST, Department Signal, February 1988, (in French).

[10] I. Csiszar, "Information-type distance measures and indirect observations", *Stud. Sci. Math. Hungar.*, Vol. 2, 1967, pp. 299-318.

[11] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems", *Ann. Probability*, Vol. 3, February 1975, pp. 146-158.

[12] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1982.

[13] D. Dacunha-Castelle, Inégalités sur les couples de probabilités, Summer School St. Flour, 1977, Chapter 3, (in French).

[14] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic work recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-28, No. 4, August 1980, pp. 357-366.

[15] P.A. Devijver, "On a new class of bounds on Bayes risk in multihypothesis pattern recognition", *IEEE Trans. Comput.* Vol. C-23, No. 1, January 1974.

[16] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.

[17] W. Gersch, "Nearest neighbor rule classification of stationary and nonstationary time series", in: D.F. Findley, ed., *Applied Time Series Analysis*, N.Y. Acacemic, 1981, pp. 221-270.

[18] A.H. Gray and J.D. Markel, "Distance measures for speech processing", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-24, No. 5, October 1976, pp. 380-391.

[19] R.M. Gray, A. Buzo, A.H. Gray and Y. Matsuyama, "Distortion measures for speech processing". *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-28, No. 4, August 1980, pp. 367-376.

[20] R.M. Gray, D.L. Neuhoff and P.C. Shields, "A generalization of Ornstein's $\bar{d}$ distance with applications to information theory", *Ann. Probability*, Vol. 1, 1975, pp. 315-328.

[21] R.M. Gray, *Probability, Random Processes, and Ergodic Properties*, Springer-Verlag, New York, 1988.

[22] U. Grenander and G. Szegö, *Toeplitz forms and their Applications*, Univ. California Press, Berkeley, 1968.

[23] Y. Grenier, "Modélisation et reconnaissance de la parole", in: *Outils et modèles mathématiques pour l'Automatique, l'Analyse des Systèmes et le Traitement du Signal*, Editions du CNRS, Vol. 2, 1982, pp. 617-637, (in French).

[24] N. Ishii, A. Iwata and N. Suzumura, "Segmentation of nonstationary time-series", *Int. J. Syst. Sci.*, Vol. 10, No. 8, August 1979, pp. 883-894.

[25] N. Ishii, H. Sugimoto, A. Iwata and N. Suzumura, "Computer classification of the EEG time-series by Kullback information measure". *Int. J. Syst. Sci.*, Vol. 11, No. 6, June 1980, pp. 677-688.

[26] F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-23, No. 1, February 1975, pp. 67-72.

[27] F. Itakura and S. Saito, "An analysis-synthesis telephony based on maximum likelihood method", *Proc. Int. Conf. Acoust.*, c-5-5, 1968, pp. c17-c20.

[28] F. Itakura and T. Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum". *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-87)*, Dallas, TX, pp. 1257-1260.

[29] R.W. Johnson, J.E. Shore and J.P. Burg, "Multisignal minimum-cross-entropy spectrum analysis with weighted initial estimates", *IEEE Trans. Acoust. Speech., Signal Process.*, Vol. ASSP-32, No. 3, June 1984, pp. 531-539.

[30] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection", *IEEE Trans. Commun.*, Vol. 15, 1967, pp. 52-60.

[31] D. Kazakos, "The Bhattacharyya distance and detection between Markov chains", *IEEE Trans. Inf. Theory*, Vol. IT-24, No. 6, November 1978, pp. 747-754.

[32] D. Kazakos and P. Papantoni-Kazakos, "Spectral distance measures between gaussian processes". *IEEE Trans. Autom. Control*, Vol. AC-25, No. 5, October 1980, pp. 950-959.

[33] J. Kittler, "Mathematical methods of feature selection in pattern recognition", *Int. J. Man-Machine Studies*, Vol. 7, 1975, pp. 609-637.

[34] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[35] S. Kullback, J.C. Keegel and J.H. Kullback, "Topics in statistical information theory", *Lecture Notes in Statistics.*, Vol. 42, Springer-Verlag, New York, 1988.

[36] K. Lashkari, B. Friedlander, J. Abel and B. McQuiston, "Classification of transient signals", *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-88)* New York, pp. 2689-2692.

[37] E.L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, 1959.

[38] J. Makhoul and A.O. Steinhardt, "On matching correlation sequences by parametric spectral models". *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-87)*, Dallas, TX., pp. 995-998.

[39] D. Mansour and B.H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition". *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-88)*, New York, NJ.

[40] Y. Matsuyama, "Mismatch robustness of linear prediction and its relationship to coding", *Inf. Control*, Vol. 47, 1980, pp. 237-262.

[41] R.K. Mehra, "Optimal input signals for parameter estimation in dynamic systems—Survey and new results". *IEEE Trans. Autom. Control*, Vol. AC-19, No. 6, December 1974, pp. 753-768.

[42] N. Nocerino, F.K. Soong, L.R. Rabiner and D.H. Klatt, "Comparative study of several distortion measures for speech recognition", *Speech Commun.*, Vol. 4, 1985, pp. 317-331.

[43] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1975.

[44] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden Day, San Francisco, 1964.

[45] H.V. Poor, "Robust decision design using a distance criterion", *IEEE Trans. Inf. Theory.* Vol. IT-26, No. 5, September 1980, pp. 575-587.

[46] C.R. Rao and T.K. Nayak, "Cross entropy, dissimilarity measures, and characterizations of quadratic entropy", *IEEE Trans. Inf. Theory*, Vol. IT-31, No. 5, September 1985, pp. 589-593.

[47] C.R. Rao, "Differential metrics in probability spaces", in: Shanti S. Gupta, ed., *Institute of Math. Stat.—Lecture Notes: Monograph Series*, Vol. 10, 1987.

[48] J.E. Shore and R.W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy", *IEEE Trans. Inf. Theory*, Vol. IT-26, No. 1, January 1980, pp. 26-37.

[49] J.E. Shore, "Minimum cross-entropy spectral analysis", *IEEE Trans. Acoust. Speech., Signal Process.*, Vol. ASSP-29, No. 2, April 1981, pp. 230-237.

[50] J.E. Shore and R.M. Gray, "Minimum cross-entropy pattern classification and cluster analysis", *IEEE Pattern*

Anal. Mach. Intell., Vol. PAMI-4, No. 1, January 1982, pp. 11-17.

[51] S. Sugimoto and T. Wada, "Spectral expressions of information measures of gaussian time series and their relation to AIC and CAT", *IEEE Trans. Inf. Theory*, Vol. IT-34, No. 4, July 1988, pp. 625-631.

[52] T. Soderstrom and K. Kumamaru, "Some model validations criteria based on Kullback discrimination index", *Proc. 24th IEEE Conf. Decision Control 85*, Fort Lauderdale, Fl., pp. 219-224.

[53] F.K. Soong and M.M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-36, No. 1, January 1988, pp. 41-48.

[54] P.V. de Souza, "Statistical tests and distance measures for LPC coefficients", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-25, No. 6, December 1977, pp. 554-559.

[55] P.V. de Souza and P.J. Thomson, "LPC distance measures and statistical tests with particular reference to the likelihood ratio", *IEEE Trans. Acoust. Speech, Signal Process.*, Vol. ASSP-30, No. 2, April 1982, pp. 304-315.

[56] J.M. Tribolet, L.R. Rabiner and M.M. Sondhi, "Statistical properties of an LPC distance measure", *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-79)*, pp. 739-743.

[57] P.M. Trouborst, E. Backer, D.E. Boekee and I.J. Boxma, "New families of probabilistic distance measures", *Proc. 2nd Int. Joint Conf. Pattern Recognition*, Copenhague, 1974.

[58] C. Villemur, F. Castanie and B. Georgel, "Modélisation paramétrique et classification automatique de signaux de forme transitoire", *Proc. GRETSI 87*, Nice, (in French).

[59] B. Yegnanarayana and D.R. Reddy, "A distance measure based on the derivative of linear prediction phase spectrum", *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP 79)*, pp. 744-747.