Distance measures for signal processing and pattern recognition

Michèle Basseville

INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 899

DISTANCE MEASURES FOR SIGNAL PROCESSING AND PATTERN RECOGNITION

Programme 5

Michèle BASSEVILLE

Septembre 1988

RR_0899

DISTANCE MEASURES FOR SIGNAL PROCESSING AND PATTERN RECOGNITION

Michèle BASSEVILLE

Publication Interne n° 422

Septembre 1988

IRISA INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

DISTANCE MEASURES FOR SIGNAL PROCESSING AND PATTERN RECOGNITION *

Michèle Basseville
IRISA/CNRS
Campus de Beaulieu
F-35042 Rennes Cedex

Publication Interne n° 422 - Septembre 1988 - 50 Pages

Abstract.- We present some general tools for measuring distances either between two statistical models or between a parametric model (or signature) and a signal. These tools are useful for solving a variety of Signal Processing problems such as detection, segmentation, classification, recognition or coding.
After a section devoted to general distance measures between probability laws, we investigate the question of spectral distances between processes. Then we describe results concerning AR and ARMA models, for which we also mention the problems related to the interaction between distances for parametric models and estimation of the parameters of these models. We also recall (when necessary) 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.

DISTANCES EN TRAITEMENT DU SIGNAL ET RECONNAISSANCE DES FORMES

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.

Running headline.- Distances for Signal Processing.
47 pages - 1 table.
Keywords.- distances, detection, classification, segmentation, recognition, coding.

TABLE OF CONTENTS

I - 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, choice of optimal input signals for system identification)... .

Up 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 neither the nature of the parameters which characterize the probability laws nor the way by which they have been estimated. On the other hand, there are a lot of specific studies in the speech processing domain (coding, recognition), where refinements of Itakura or cepstral distance measures still emerge now.

The aim of this paper is to get together disseminated tools and results concerning distance measures, in view of application in Signal Processing, for detection and recognition in general. Especially, we shall address some typical issues in model based Signal Processing, namely choice of models, parametrizations and parametric estimation methods on one hand, and choice of distance measures between these models on the other one, without forgetting the possible interaction between these two choices.

However we do not claim that we exhaustively compiled all the litterature concerning distance measures. Nevertheless, we try to follow a presentation going from a general framework to particular cases.

In section II, 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 III is devoted to spectral distance measures between processes. In section IV, we analyse the results related to AR or ARMA models and to the interaction parametrization/distance.

Let us emphasize that the word distance here means 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 don't all satisfy the triangular inequality.

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

- $d(P_1, P_2)$ distance between the probability laws P_1.et P_2
  $d(A_1, A_2)$ distance between the parametric models A_1 et A_2

- $d(y_1, A_2)$ distance between a signal (y_1) and a model A_2

- with such notations, if $\hat{A}_i$ represents an estimate of A_i (i = 1,2), then:

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

are distances between signals. the symbol ^ will often be omitted for simplification.

II - GENERAL DISTANCE MEASURES

In this section, we introduce general classes of distance measures, or divergence coefficients, between probability distributions. In II.1, we start with the class related to Csiszar f-divergence (9) which contains many known distance measures which we also recall in the multidimensional case. But this class is not related to information measures, except for Kullback divergence. Then in II.2 we describe the so called class of general mean distance introduced by Boeker and Van der Lubbe [6] for Pattern Recognition, which is directly related to information measures. In section II.3 we investigate a general contrast criterion which may be used as a distance and was introduced by Poor [41] for robust detection. Then we describe some general tools for measuring the distance between a model and observations: in section II.4, we recall the axiomatic derivation of the entropy principle due to Shore [42], and finally in section II.5 we present a general model validation tool to be used for segmentation or monitoring [4].

II.1 - f-DIVERGENCE

This general notion has been apparently introduced by Csiszar [9] [10] and independently by Ali et Silvey [1]. It is based upon the fact that it is intuitively "natural" to measure the remoteness of two probability distributions p_1 et p_2 with the aid of the "dispersion" - with respect to p_1 - of the likelihood ratio $\phi(x) = \frac{p_2(x)}{p_1(x)}$: if p_1 and p_2 are two densities on $\mathbb{R}$, when they "move" away from each other, ϕ increases on a set of decreasing p_1-probability and decreases on a set of increasing p_1-probability. More generally, we get "reasonable" divergence coefficients by considering as a dispersion measure of ϕ the p_1-expectation of any increasing function g of this p_1-expectation.

II.1.1 - Definition

More precisely, let f a continuous convex real function on $\mathbb{R}_+$ (weaker conditions than continuity may be found in [9] [6]), and let g be an increasing function on $\mathbb{R}$. Consider the following class of divergence coefficients between two probability laws P_1 et P_2 over the same space:

$$d(P_1, P_2) = g\left[\mathbb{E}_1\left[f\left(\frac{dP_2}{dP_1}\right)\right]\right] \tag{1}$$

where
$$\frac{dp_2}{dp_1} \triangleq \phi \tag{2}$$

is the Radon-Nikodym derivative (possibly generalized in the case where P_2 has a singular component with respect to P_1 see [1]), and where $\mathbb{E}_1$ is the expectation with respect to P_1.

II.1.2. Properties

Then [1] d has the following properties :

i) if $y = t(x)$ is a measurable transformation of $(\mathcal{X},\mathcal{F})$ on $(\mathcal{Y},\mathcal{G})$ 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 by increasing the number of observations, i.e.:

$$d(P_1^{(m)}, P_2^{(m)}) \leq d(P_1^{(n)}, P_2^{(n)}) \qquad \text{for } m < n \qquad (3')$$

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

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

iii)if $(p_\theta ; \theta \in]\, a, b\, [)$ is a family of densities on $\mathbb{R}$ with monotone likelihood ratio

(ie if there exists a function T such that for any $\theta_1 < \theta_2$, $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$ is an increasing function of T(x) ; [33], p68 ; important condition for designing tests-), then for $a < \theta_1 < \theta_2 < \theta_3 < b$, we have:

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

Let us notice that the convexity of f is a necessary condition for i).Furthermore, for g identity and p_1, p_2 two densities, we have [10] :

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

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. 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 formulas (18) to (21) -gave rise to many studies about distance measures [6] [7] [13], 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].

II.1.3. Examples

Let λ be a measure on $(\mathcal{X}, \mathcal{F})$ such that P_1 and P_2 are absolutely continuous with respect to λ, with densities p_1 and p_2 (ex : $\lambda = P_1 + P_2$ or Lebesgue measure).

- **Kolmogorov variational distance**

$$f(x) = |1 - x| \; ; \; g(x) = \frac{x}{2} ;$$

$$d(P_1, P_2) = \frac{1}{2} \int_{\mathcal{X}} |p_2 - p_1| \, d\lambda \triangleq V(P_1, P_2) \qquad (5)$$

- **Hellinger distance**

$$f(x) = (\sqrt{x} - 1)^2 \; ; \; g(x) = \frac{x}{2} \; ;$$

$$d(P_1, P_2) = \frac{1}{2} \int_{\mathcal{X}} (\sqrt{p}_2 - \sqrt{p}_1)^2 \, d\lambda \triangleq H^2(P_1, P_2) \qquad (6)$$

- **Kullback information**

$$f(x) = - \text{Log } x \; ; \; g(x) = x \; ;$$

$$d(P_1, P_2) = \int_{\mathcal{X}} p_1 \text{ Log } \frac{p_1}{p_2} \, d\lambda \triangleq K(P_1, P_2) \qquad (7)$$

• **Kullback divergence**

$f(x) = (x - 1) \operatorname{Log} x \; ; \; g(x) = x \; ;$

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

which is symmetrical.

• **Chernoff distance**

$$0 \leq r \leq 1 \; ; \; f(x) = -x^{1-r} \; ; \; g(x) = - \operatorname{Log} (-x) \; ;$$

$$d(P_1, P_2) = - \operatorname{Log} C(P_1, P_2)$$

where $$C(P_1, P_2) = \int_{\mathcal{X}} p_1^r \, p_2^{1-r} \, d\lambda \qquad (9)$$

is called Chernoff coefficient.

• **Bhattacharyya distance** :previous case with $r = \frac{1}{2}$

ie $$f(x) = - \sqrt{x} \; ; \; g(x) = - \operatorname{Log} (-x) \; ;$$

$$d(P_1, P_2) = - \operatorname{Log} \rho (P_1, P_2) \triangleq B(P_1, P_2)$$

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

is called Bhattacharyya coefficient in the field of Pattern Recognition and affinity in theoretical Statistics. We refer to [28] for its formulation in the case of Markov chains and its use for detection .

• **Generalized Matusita distance**

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

$$d(P_1, P_2) = \sqrt[r]{\int_{\mathcal{X}} | p_1^{1/r} - p_2^{1/r} |^{r} \, d\lambda} \triangleq M_r (P_1, P_2) \qquad (11)$$

Notice that, for r = 1, we get Kolmogorov 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 et $1-\pi$ and where the corresponding densities of the observations are p_1 and p_2, is :

$$P_e = \int_{\mathcal{X}} \min \left[\pi p_1, (1 - \pi) p_2 \right] d\lambda \qquad (12)$$

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 [31] **Patrick and Fisher** distance :

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

and **Lissack and Fu distance**

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

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 III.

• **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 [14] [31] :

* Bhattacharyya distance:

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

* Kullback divergence

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

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

* **Mahalanobis distance:**

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

II.1.4 - Some inequalities

As we said before, the search for bounds of the classification error probability [7] [13] [5] [6], but also other goals such as feature selection for Pattern Recognition [7] [31] or signal selection [27] [37],led to various inequalities between P_e and many of the above mentioned distance measures or between these distances.

For example [27] [30] :

$$\frac{1}{2}\left[1 - \sqrt{1 - 4\pi(1-\pi)\rho^2}\right] \leq P_e \leq \sqrt{\pi(1-\pi)}\ \rho \quad (18)$$

where ρ is defined in (10)

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

or [6] :

$$P_e \leq \frac{1}{2} - \frac{1}{2} V \quad (20)$$

$$P_e \leq \frac{1}{2} - \frac{1}{2} M_r^r \quad (21)$$

Other bounds for P_e may be found in [13] [7], [6] where the case of several classes is also investigated and general bounds are given , and [5]. In [29] the author studies the case where the a priori probability laws are not precisely known .

Among the known theoretical inequalities[11]., we have :

$$H^2(2 - H^2) = 1 - \rho^2 \tag{22}$$

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

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

$$\frac{1}{4}\, e^{K(P_1,\, P_2)} \leq 1 - V(P_1, P_2) \leq \rho\,(P_1, P_2) \tag{25}$$

II.2 - General mean distance for classification

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

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

$$= 1 - \int_{\mathcal{X}} p(x) \left[\max_i\; P(C_i \mid x)\right] dx$$

where $P(C_i \mid x)$ is the a posteriori probability of the class C_i given the observation x,

and $p(x) = \sum_{i=1}^{m} \pi_i\; P(x \mid C_i)$.

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 :

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

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 .

Examples

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

$$\sum_i - P(C_i \mid x) \ \text{Log}\, P(C_i \mid x)$$

- **Devijver Bayesian distance[13]**

  * $\beta = 2$, $\alpha = 1$;

$$G_{1,2}(C \mid X) = \int_{\mathcal{X}} p(x) \left[\sum_{i=1}^{m} P(C_i \mid x)^2 \right] dx \qquad (27)$$

$$\triangleq B(C \mid X)$$

$$= 1 - H_2(C \mid X)$$

where H_2 is the mean conditional quadratic entropy defined from the usual entropy by replacing $- \log P(C_i \mid x)$ by $1 - P(C_i \mid x)$.

* $\beta = 3$, $\alpha = 1$

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

It can be shown that :

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

where H_3 is the mean conditional cubic entropy introduced by Chen [7] and defined from the usual entropy by replacing $\log P(C_i \mid x)$ by

$$P(C_i \mid x) - 1 + \frac{1}{2}\left[P(C_i \mid x) - 1\right]^2.$$

* $\alpha = 1/\beta$ and $\beta > 1$:

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

is the distance $B'_R(C \mid X)$ proposed by Trouborst [50].

Many bounds for P_e can be obtained from this class. Among others[6] :

* $\alpha > 0, \beta > 1, \ 1 \le \alpha\beta < 1 + \alpha \Rightarrow$

$$1 - G_{\alpha,\beta}(C \mid X)^{1/\alpha\beta} \le P_e \le 1 - G_{\alpha,\beta}(C \mid X)^{1/\alpha(\beta-1)}$$

$$\le 1 - \frac{1}{m^{\alpha}} G_{\alpha,\beta}(C \mid X)$$

* $\alpha > 0, \beta > 1, \alpha\beta \le 1 \Rightarrow$

$$1 - G_{\alpha,\beta}(C \mid X) \le P_e \le 1 - G_{\alpha,\beta}(C \mid X)^{1/\alpha(\beta-1)}$$

$$\le 1 - \frac{1}{m^{1/\beta}} G_{\alpha,\beta}(C \mid X)^{1/\alpha\beta}$$

* $\alpha > 0, \beta > 1, \alpha\beta \ge \alpha + 1 \Rightarrow$

$$1 - G_{\alpha,\beta}(C \mid X)^{1/\alpha\beta} \le P_e \le 1 - G_{\alpha,\beta}(C \mid X)$$

$$\le 1 - \frac{1}{m^{\alpha}} G_{\alpha,\beta}(C \mid X)$$

* $$\lim_{\beta \to \infty} (1 - G_{1/\beta,\beta}(C \mid X)) = P_e$$

These are generalizations of known bounds.

II.3 - CONTRAST TYPE DISTANCE MEASURES

Another type of distance between laws has been introduced by Poor [41] for robust detection. It is based upon a generalized version of the signal to noise ratio often called contrast .

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

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

If P_1 and P_2 have densities p_1 et 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 $\phi = \dfrac{p_2}{p_1}$. 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 $\quad f(x) = (x - 1)^2,\ g(x) = x \tag{32}$

The interest 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) reduces to the derivation of a least favorable pair in terms of risk - LMFR in abbreviated form - ; the risk robust detector is then the likelihood ratio of this LMFR pair. The problem is that the finding of 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 [41] that, if we define a robustness notion in terms of the distance S (29), we keep the fact that the robust detector in terms of S - LMFS in abbreviated form - ; but we gain 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 :

$$S_h(P_1, P_2) = \frac{(\mathbb{E}_2(h))^2}{\mathbb{E}_1(h^2)} \qquad (33)$$

also used for detection.

Poor also shows [41] 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, refering to section II.5 for the local point of view, $Var_1(\phi) = S_\phi(P_1, P_2)$ plays the same role as Fisher information $I(p) = \int \frac{(p')^2}{p}$ 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 II is devoted to distance measures not between laws but between a law (or a model) and datas. This classification is somewhat arbitrary, because we would have have delt 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.

Paragraphs 4 and 5 are even less claimed to be exhaustive than the previous ones. The presented tools will be re-analysed in section IV devoted to parametric AR and ARMA models.

II.4 - ENTROPY

In this section, we give the axiomatic derivation of the maximum entropy and minimum cross-entropy (or divergence) principles due to Shore and Johnson [42], because it emphasizes the criteria which lead to these distance measures between models and data already introduced by Kullback [32].

Given a system with the following informations :

- an a priori density p ;
- constraints I on the "true" unknown density q* of the form :

$$\int q^*(x)\ a_k(x)\ dx = 0$$

or (34)

$$\int q^*(x)\ c_k(x)\ dx \geq 0$$

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 (34).
We define 4 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 cross-entropy (or "oriented" divergence or Kullback information (7)) :

$$K(q,p) = \int_{\mathcal{X}} q(x)\ \mathrm{Log}\frac{q(x)}{p(x)}\ dx \qquad (35)$$

For this purpose, we introduce the following "information operator" 0 :

$$q = p\ 0\ I$$

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, ie :

$$q = p\ 0\ I \iff H(q,p) = \min_{q' \text{ satisfying } I} H(q',\ p)$$

If there exists another functional H' such that

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

H' et H are said to be equivalent, and the operator 0 can be realized using either functional.

The axioms are as follows :

i) **unicity** : for any p and any I, q = p 0 I is unique ;

ii) **invariance** by coordinate transformation : if Γis a transformation from $\mathcal{X}$ to $\mathcal{Y}$ then :

$$(\Gamma_p) \text{ o } (\Gamma I) = \Gamma(P \text{ o } I)$$

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

iii) **system independence** :

If $\mathcal{X}_1$ and $\mathcal{X}_2$ 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 :

$$(p_1 p_2)\ 0\ (I_1 \wedge I_2) = (p_1 0\ I_1)\ (p_2\ 0\ I_2) \tag{37}$$

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.

iv) **subset independence**

If $\mathcal{X}$ is an 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:

$$(p\ 0\ I) * S_i = (p * S_i)\ 0\ I_i \tag{38}$$

where $I = I_1 \wedge \ldots \wedge I_n$.

(In fact, a stronger condition is imposed [42]).

In order to show that an operator 0 satisfying these 4 axioms can be realized only by the cross-entropy K (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 *iv)* (38) and a special case of the axiom *ii)* (36) lead to restricted functionals of the form :

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

Then, at the second step, the general case of *ii)* is shown to lead to the form :

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

(which strangely enough looks like the remark below (4)) !.

The third step uses axiom *iii)* and shows that, if H satisfies the 4 axioms, H is equivalent to the cross-entropy K (35). The last step shows that K actually satisfies the 4 axioms.

This cross-entropy minimization principle is successfully used for spectral analysis [43] including in the multidimensional case [26], classification for Pattern Recognition [44] and many other applications in various domains (see [42]).

II.5 - MODEL VALIDATION

We conclude this section with another tool for measuring the distance between a model and datas, introduced for signal segmentation and systems monitoring [4]. An example of use of this measure will be presented in section IV devoted to the parametric models ARMA.

Let (Y_n) be a controlled Markovian process (or more generally a controlled semi-Markovian 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. it 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 fixed by the user, and let us investigate the problem of detecting small deviations (local approach as at the end of II.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 :

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

where

$$R(\theta_0) = \sum_{n \in Z} \text{cov}\,(H(\theta_0\, Y_n) \;,\; H(\theta_0, Y_0))$$

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

(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 Markovian 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 [45] for a special use of Kullback information (7) between condotional laws for multivariable input/output model validation.

Before entering the section devoted to spectral distance measures, let us notice the existence - and the huge theoretical importance - of a general distance between processes, called $\bar{\rho}$ Ornstein distance [18], which 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. An example concerning gaussian processes will be seen in the paragraph III.3.5.

III - 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 the papers [16] [17] [36] [30], and also [18] and the book [40] which are less accessible.

III.1 - PRELIMINARY REMARKS

Following [17], we shall use the following notations. Let $s(\lambda)$ be a (energy or power) spectral density corresponding to a scalar signal. λ varies from $-\pi$ à 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 :

$$s(\lambda) = \sum_{n \in \mathbb{Z}} r(n)\ e^{-jn\lambda}$$

$$r(n) = \int_{-\pi}^{\pi} s(\lambda)\ e^{jn\lambda}\ \frac{d\lambda}{2\pi} \qquad (44)$$

For a wide sense stationary ergodic process $(y_n)_{0 \le n \le N-1}$, the sequence r(n) defined by :

$$r(n) = \begin{cases} r_1^{(N)}(n) & \text{pour } |n| < N \\ 0 & \text{pour } |n| \ge N \end{cases}$$

where

$$r_1^{(N)}(n) \triangleq \sum_{k=0}^{N-|n|-1} y_k \, y_{k+|n|} \qquad (0 \le |n| \le N-1)$$

corresponds to an energy spectral density.

If $r(n) = r_2(n) \triangleq \mathbb{E}(y_k y_{k+n})$, then (44) defines a power spectral density. Noting that for any n :

$$\lim_{N \to \infty} \frac{1}{N} r_1^{(N)}(n) = r_2(n) \quad \text{p.s.},$$

we conclude that the sequel will not depend upon the nature - energy or power - of the spectral density .
Following the terminology introduced in the introduction, for $r = r_2$ the following distance measures will be laws between processes, and for $r = r_1$ (empirical covariance) distances between signals.

Let $R_N(s)$ be the Toeplitz $(N+1) \times (N+1)$ matrix , the (k,j)th element of which is $r(k-j)$ $(0 \le k, j \le N)$. We shall use several fundamental properties of R_N [19] [17] $|R_N|$ denotes the determinant of R_N.

For each p, there is associated with the spectral density s a Toeplitz form :

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

$$= \sum_{k=0}^{p} \sum_{l=0}^{p} a_k a_l r(k-l) \qquad (45)$$

$$= a^T R_p(s) a$$

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

A numerically convenient form is :

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

where

$$r_a(k) \triangleq \sum_{l=0}^{p-k} a_l a_{l+k} \qquad (0 \le k \le p)$$

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

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

Let be $$\sigma_s^2(p) = \min_{\substack{a \\ (a_0 = 1)}} T_p(a) \quad ; \quad \text{then [19] :}$$

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

and the minimizing polynomial A(z) may be analytically expressed in terms of orthogonal polynomials (cf. Levinson algorithm). Let $A_p(z)$ be the p th 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 :

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

we can write :

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

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

Furthermore, let be [19] :

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

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

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 $\frac{\sigma_s^2}{A}$ the (infinite) autoregressive model of s, and $\frac{1}{A}$ the normalized AR model.

Most of the spectral distance measures which we shall consider will be in terms of L_q norms, ie :

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

which satisfies :

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

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

III.2 - SPECTRAL DISTANCE MEASURES AND EQUIVALENCES

Spectral distances between two spectral densities s_1 et s_2 may be measured with the aid of L_q norms of their difference, ie :

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

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 II.1. However, the spectral distances which will be used here are functions of the **difference between the log-spectra**, ie of the ratio between the spectra :

$$d(s_1, s_2) = d(1, \frac{s_2}{s_1}) = d(\frac{s_1}{s_2}, 1) \qquad (50)$$

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

For a given distance d, we shall use two types of scaling [17]. A **gain normalized distance measure** is defined by:

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

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

A **gain optimized distance measure** is defined by :

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

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 numbers of 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 [17], 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), ie 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 :

$$d_1 <\equiv> d_2 .$$

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'.**

III.3 - MAIN SPECTRAL DISTANCE MEASURES

III.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 :

$$\begin{aligned} d_q(s_1, s_2) &= \| \text{Log } s_1 - \text{Log } s_2 \|_q \\ &= \| \text{Log} \frac{s_1}{s_2} \|_q \end{aligned} \qquad (53)$$

The more common choices are :

- $q = 1$ mean absolute distance
- $q = 2$ mean quadratic distance (r m s)
- $q = \infty$ 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 $\frac{10}{\text{Log}(10)} = 4.34$. 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_∞ [16], at least when the spectra are estimated via Fourier transform.

III.3.2 - Itakura-Saïto distance [24]

It is defined by :

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

and is also called "error matching measure".

As : $u - \log u - 1 \geq 0$, we also have :

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

by the residual theorem.

Using the expansion :

$$u = \exp(\text{Log } u) = 1 + \text{Log } u + \frac{1}{2} (\text{Log } u)^2 + \ldots$$

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 :

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

ie for given spectral gains, constant spectra give the smallest distorsion.

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 - namely if we want to solve the problem of linear prediction of s_1 - from (55) and (45) we conclude that :

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

We shall hark 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 [40] :

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

It can be shown that [40] :

$$\begin{aligned} K(s_1, s_2) &\triangleq \lim_N \frac{1}{N} K_N(s_1, s_2) \\ &= \frac{1}{2} d_{IS}(s_1, s_2) \end{aligned} \qquad (58)$$

In other words, Itakura-Saïto distance is equal to two times tha asymptotical Kullback information under gaussian hypothesis. This technique has been successfully tested for classifying non gaussian data for the purpose of recognition of EEG signals [15]. 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 [17]. This is also the case for classification [22] and recognition of EEG signals once more, for which in [21] Kullback distance, Kullback divergence and Bhattacharrya distance have been compared.

III.3.3. Itakura distance

$$\begin{aligned} d_I(s_1, s_2) &\triangleq d_{IS}(s_1, s_2) \\ &= \min_{\alpha \geq 0} d_{IS}(s_1, \alpha s_2) \end{aligned} \qquad (59)$$

From (55), we get :

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

and

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

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

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

This distance is also called log likelihood ratio because of its asymptotical expression in the gaussian case[23] [48]. We refer to the next section for additional details.

III.3.4 - Model distance measure

Also introduced by Itakura [23], it is defined by :

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

where A_1 and A_2 are the normalized AR models for s_1 and s_2. It can be shown that [17] :

$$d_m^*(s_1, s_2) = \left\| \frac{A_2}{A_1} \right\|_2^2 - 1$$

$$= \exp(d_I(s_1, s_2)) - 1 \qquad (64)$$

and thus d_m^* and d_I are completely equivalent. We also have :

$$d_m^* = d_{IS}^*$$

This distance was introduced as an approximation for d_I for d_I small (cf. (64)). It is always an upper bound for d_I.

It is called a model distance measure because it measures how nearly the normalized models or filters A_1 and A_2 are to being inverses (see next section).

A similar unnormalized model distance is given by :

$$d_m(s_1, s_2) = \left\| 1 - \frac{\sigma_1 / A_1}{\sigma_2 / A_2} \right\|_2^2$$

$$= \frac{\sigma_1^2}{\sigma_2^2} d_m^*(s_1, s_2) + \left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^2$$

$$= \frac{\sigma_1^2}{\sigma_2^2} d_m^*(s_1, s_2) + d_m(\sigma_1^2, \sigma_2^2)$$

But :
$$d_{IS}(s_1, s_2) = \frac{\sigma_1^2}{\sigma_2^2} d_m^*(s_1, s_2) + d_{IS}(\sigma_1^2, \sigma_2^2)$$

thus : $d_{IS} \Leftrightarrow d_m$

However d_{IS} and d_m are not NN-equivalent.

The optimization of d_m according to (52) gives :

$$d'_m(s_1, s_2) = 1 - \frac{1}{\left\| \frac{A_1}{A_2} \right\|_2^2} \qquad (66)$$

which can be shown to be a monotonic function of d_m^* , and thus :

$$d'_m \Longleftrightarrow d_m^*$$

III.3.5 - Symmetrized distance measures

A spectral distance measure d can be symmetrized by considering arithmetical or geometricel means of $d(s_1, s_2)$ and $d(s_2, s_1)$, namely by defining for $q \geq 1$:

$$d^{(q)}(s_1, s_2) = \frac{1}{2}\left(d(s_1, s_2)^q + d(s_2, s_1)^q\right)^{1/q} \qquad (67)$$

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

A symmetrized version of Itakura-Saito distance was introduced in [16] and defined by

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

where the terminology cosh (of the spectral difference measured on a logarithmic scale, ie $\mathrm{Log}\frac{s_1}{s_2}$) comes from (54).

d_{cosh} is related to a decibel scale [16] with the aid of the quantity D such that :

$$\cosh(D) - 1 = d_{cosh}$$

namely :

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

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 [18] [36]

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

where $Y^{(1)}$ and $Y^{(2)}$ are two gaussian processes with spectral densities $\frac{s_1}{s_2}$ et $\frac{s_2}{s_1}$

respectively, and where $\bar{\rho}$ is the Ornstein distance [18] between processes already mentioned at the end of section I. For gaussian processes $X^{(1)}$ et $X^{(2)}$ with spectral densities s_1 and s_2, we have

$$\bar{\rho}(X^{(1)}, X^{(2)}) = 2\, H^2(s_1, s_2)$$

where H^2 is Hellinger distance defined in (6). This leads to the following (simple) relationship:

$$d_{cosh}(s_1, s_2) = H^2\left(\frac{s_1}{s_2}, \frac{s_2}{s_1}\right) \tag{70}$$

III.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 symmetrical distances.
Recall that we always have :

$$d^{(1)} \Rightarrow d$$

$$d^{*} \Rightarrow d'$$

$$d \Rightarrow d'$$

The known equivalences between the above distance measures are summarized in the following diagram [17] :

$$
\begin{array}{ccccccccc}
d_I^{(1)} & \Leftrightarrow & d_{cosh} & \Leftrightarrow & d_m^{(1)*} & \Rightarrow & d_I = d_{IS} & \Leftrightarrow & d_m^* = d_{IS}^* \Leftrightarrow d_m \\
 & & \Uparrow & & & & \Uparrow & & \\
d_{cosh} = & & d_{IS}^{(1)} & \Leftrightarrow & d_m^{(1)} & \Rightarrow & d_{IS} & \Leftrightarrow & d_m \\
 & & \Downarrow & & & & & & \\
 & & d_2 & & & & & &
\end{array}
$$

Other results are described in [36] together with their consequences on robustness issues of linear predictive coding.

III.3.7 - The case of multidimensional gaussian processes

In [30] 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 diffference between the process means.
For example [30] :

$$
2\,K(Y^{(2)}, Y^{(1)}) = \int_{-\pi}^{\pi} \left(tr\, S_1^{-1}(\lambda) \left[S_2(\lambda) - S_1(\lambda) \right] - \log S_1^{-1}(\lambda)\, S_2(\lambda) \right) \frac{d\lambda}{2\pi}
$$

$$
+ \int_{-\pi}^{\pi} \sum_{k=1}^{r} \sum_{j=1}^{r} m_{kj}(\lambda)\; s_{kj}(0,\lambda) \frac{d\lambda}{2\pi}
$$

where

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

$$
\left[(1 - t)\, S_1(\lambda) + t\, S_2(\lambda) \right]^{-1} = (s_{kj}(t,\lambda))_{1 \le k,j \le r}
$$

IV. 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.

IV.1 - L_2-NORM AND CEPSTRAL DISTANCE

In the last section (III.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 cercle, we define the **cepstral coefficients** [16] by the coefficients of the Taylor expansion of the logarithm of the filter transfer function, ie :

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

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

$$c_0 = \mathrm{Log}(\sigma^2) \qquad (74)$$

$$c_{-k} = c_k$$

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

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

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

$$c_n = \sum_{k=1}^{n-1} (1 - \frac{k}{n})\, h_k\, c_{n-k} + h_n \quad (n > 1)$$
$$c_1 = h_1 \qquad (76)$$

or from the linear prediction coefficients by :

$$c_1 = - a_1$$
$$c_n = - a_n - \sum_{k=1}^{n-1} \frac{k}{n}\, c_k\, a_{n-k} \qquad 1 < n \leq p \qquad (77)$$
$$c_n = - \sum_{k=1}^{p} \frac{n-k}{n}\, c_{n-k}\, a_k \qquad p+1 \leq n$$

In order to obtain (77) - or (76) -, derive the two handsides of (72) with respect to z^{-1} and use (75).

Notice that the two resulting cepstra ("Fourier" or "parametric") are not identical [2]. The difference is due to the truncation of the signal over a finite time interval which is done in a Fourier analysis and which produces a signal with poles only, whatever the content (poles or zeroes) of the transfer function is. On the other hand, the "parametric" cepstrum explicitely takes into account the hypothesis that the transfer function has only poles. The consequences of these two choices on speech recognition systems will be described later.

More generally, the cepstrum of a minimum phase rational transfer function can be defined [39] . In this case, the cepstral coefficients can be interestingly expressed as a function of the poles $(z_k)_{1 \le k \le p}$ and the zeroes $(w_k)_{1 \le k \le q}$ [39] [34] :

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

This formula is obtained from (72) by a residual calculus. A consequence will be mentioned later.

The interest of the cepstral coefficients c_n for computing the distance d_2 (53) is as follows.

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 :

$$d^2(L) = \sum_{k=-L}^{L} (c_k^{(1)} - c_k^{(2)})^2 \qquad (L \ge p) \tag{89}$$

can be shown to be positive definite and to converge, when $L \to \infty$, towards d_2^2. Moreover, experiments with speech signals [16] 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] [51]. We shall discuss further this point in the paragraph IV.4.

Moreover, from (78) and (80), we conclude that, for causal spectra :

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

where $z_i^{(j)}$ $(j = 1,2 \ , \ i = 1,p)$ are the poles of the spectrum s_j. This shows that one can be very far from a true spectral distance when one tries -in an intituively "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].

IV.2 - DISTANCES d_{IS} AND d_I

We now consider the parametric formulation of the Itakura-Saïto distance d_{IS} (54) and Itakura distance d_I (60) ; then we present a variant of d_I and we describe the link between d_I and the model validation tool introduced in I.5.

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

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

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 [46]. The weighting is done with the aid of Atal perceptual filter, which gives higher weights to spectral deviations around peaks than around valleys of the spectrum, and these weights are adapted according to an estimated signal to noise ratio. More precisely, the distance is (compare with (62)) :

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

where $A_1'(e^{j\lambda}) = A_1(\alpha e^{j\lambda})$

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

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) [35].

Finally, let us conclude this paragraph with a link between d_I (82) and the model validation tool introduced in I.5. This tool, initially designed for the validation of the AR part of a multidimensional ARMA process, looks like (82) in the special case of a scalar AR(p) process. Indeed, keeping the index 1 for the signal and 2 for the model,

define : $\theta_2^T = (a_1 \ldots a_p)$ ie $a_2^T = (1 \ \theta_2^T)$. (41) becomes :

$$Z_k = \phi_k (y_k + \phi_k^T \theta_2)$$

where $\phi_k^T = (y_{k-1}, \ldots, y_{k-p})$. We then deduce :

$$\sum_k Z_k = (\gamma_p^{(1)} \mid R_{p-1}^{(1)})\, a_2$$

where $\gamma_p^{(1)T} = (r_1^{(1)} \ldots r_p^{(1)})$, $R_{p-1}^{(1)}$ is the Toeplitz matrix used in (45), and where the covariances $r_j^{(1)}$ are the **empirical** covariances.

On the other hand, in the **local** framework of **small** deviation and with the notations of I.5., one can assume that :

$$R(\theta_2) = \sigma_1^2 R_{p-1}^{(1)}$$

(42) then becomes :

$$t = \frac{a_2^T (\gamma_p^{(1)} \mid R_{p-1}^{(1)})^T R_{p-1}^{(1)^{-1}} (\gamma_p^{(1)} \mid R_{p-1}^{(1)})\, a_2}{\sigma_1^2}$$

which reduces to :

$$t = \frac{a_2^T \tilde{R}_p^{(1)} a_2}{\sigma_1^2} \qquad (84)$$

where $$\tilde{R}_p^{(1)} = \left[\begin{array}{c|c} \sigma_1^2 & \gamma_p^T \\ \hline \gamma_p & R_{p-1}^{(1)} \end{array}\right]$$

differs from the Toeplitz matrix $R_p^{(1)}$ only via the first coefficient which is equal to $\sigma_p^{(1)}$ and not r_0.

IV.3 - SOME OTHER DISTANCE MEASURES

IV.3.1 - Variants of the cepstral distance

Let us first consider variants of the cepstral distance d_2 introduced in (79). [52] [25] [35] introduced several distance measures based upon the derivative of the phase spectrum - namely the group delay - rather than the logarithm of the spectrum which, at least for speech signals, propagates over several formants the delay which may exist only on one formant [52].

Let us consider the Taylor expansion of the phase $\phi(z)$ of Log A(z). From (72) :

$$\phi(e^{-j\lambda}) = \sum_{k=1}^{\infty} c_k \sin(k\lambda) \qquad (85)$$

Thus the expansion of the group delay is :

$$\phi'(\lambda) = \frac{d\phi\,(e^{-j\lambda})}{d\lambda} = \sum_{k=1}^{\infty} k\ c_k \cos(k\lambda) \qquad (86)$$

Introducing as a new spectral distance the quantity :

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

Yegnanarayana [52] suggests to use the following euclidian distance :

$$\tilde{d}^2(L) \sum_{k=1}^{L} k^2 \left(c_k^{(1)} - c_k^{(2)}\right)^2 \qquad (86)$$

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

More generally, in [25] Itakura recently suggested to use an euclidian distance based upon a "smoothed" group delay, namely upon $w_k\ c_k$, where

$$w_k = k^s\ e^{-k^2/2\tau^2} \qquad (s \geq 0)$$

The coefficient k^s is used for isolating the spectral peaks but also for equalizing the spectral envelop. The term $e^{-k^2/2\tau^2}$ is used for cancelling the cepstral components with high order k.

Finally, still more recently [35] introduces spectral distances based upon the cosine of the angle between two cepstral coefficients arrays - with c_0 excluded -, which are more robust than the euclidian distance between these two vectors with respect to the presence of noise :

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

and

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

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

C_j is the vector of the cepstral coefficients $c_k^{(j)}$ $(k > 0)$ $(j = 1,2)$; and where $j = 1$ is again the "test" and $j = 2$ the "reference".

Other weightings, by the inverse of the variance of the c_k computed during the learning phase, have been investigated in [8].

IV.3.2 - A divergence between conditional laws

As an end point for this incomplete catalogue, let us mention a special distance used for signal segmentation [3]. This "distance" is based upon Kullback divergence between the **conditional laws** $p_j\ (y_n \mid y_{n-1}, \ldots, y_{n-p})$ 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 $\sigma_n^{(j)}$ are approximated least squares than when they are exact. We have no theoretical explanation for this stange parameters/distance interaction (see also the remark at the end of IV.1.).

IV.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 [21] [51]. 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 conforted for example in [20] and [12], 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_I. (Notice that they cannot be compared in the table of II.3.6). Similar conclusions concerning the cepstrum have been obtained in the recent work [8].

Other comparisons have been done in [38], with different weighting variants introduced for speech signals ("spectral slope",...).

Recall that the variants of d_I or d_2 introduced respectively in [46] and [25] [35] have been compared to the original distance d_I 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 [36] [17] and is partly summarized in the table of paragraph III.3.6.

V. CONCLUSION

From this 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, from (10), (11) for r = 2, (17), (58), (69) and (70), we conclude that Kullback divergence J (8) and Hellinger distance H (6) take a key part for proving complex theoretical results as well as solving applied problems. Furthermore, it is quite stimulating to find out that the same tools are preferred by theoreticians and practicians.

In addition to J and H, we also recommend to use d_2 (79) in practice, because of its euclidian nature.

REFERENCES

[1] **S.M. ALI, D. SILVEY** : "A general class of coefficients of divergence of one distribution from another". Jal of Royal Statistical Society, B, vol. 28, n°1, p.131-142, 1966.

[2] **B.S. ATAL** : "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". Jal of the Acoustical Society of America, vol. 55, n°6, p.1304-1312, June 1974.

[3] **M. BASSEVILLE** : "The two-models approach for the on-line detection of changes in AR processes". dans "*Detection of abrupt changes in signals and dynamical systems*", ed. M. Basseville, A. Benveniste, LNCIS n°77, Springer-Verlag, 1986.

[4] **A. BENVENISTE, M. BASSEVILLE, G. MOUSTAKIDES** : "The asymptotic local approach to change detection and model validation". IEEE Trans. Automatic Control, vol. AC-32, n°7, p. 583-592, July 1987.

[5] **D.E. BOEKEE, J.C. RUITENBEEK** : "A class of lower bounds on the Bayesian probability of error". Information Sciences, vol. 25, p. 21-25, 1981.

[6] **D.E. BOEKEE, J.C.A. VAN DER LUBBE** : "Some aspects of error bounds in feature selection". Pattern Recognition, vol. 11, p. 353-360, 1979.

[7] **C.H. CHEN** : "On information and distance measures, error bounds, and feature selection". Information Sciences, vol. 10, p. 159-173, 1976.

[8] **J.P. CORDEAU** : "Un système de reconnaissance - Analyse de quelques métriques". Stage Report, ENST, Department Signal, February 1988. In French.

[9] **I. CSISZAR** : "Information-type distance measures and indirect observations". Stud. Sci. Math. Hungar, vol.2, p. 299-318, 1967.

[10] **I. CSISZAR** : "I-divergence geometry of probability distributions and minimization problems". Annals of Probability, vol. 3, p. 146-158, Feb. 1975.

[11] **D. DACUNHA-CASTELLE** : "Inégalités sur les couples de probabilités". Summer School St Flour, 1977, Chapter 3. In French.

[12] **S.B. DAVIS, P. MERMELSTEIN** : "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-28, n°4, p. 357-366, Aug. 1980.

[13] **P.A. DEVIJVER** : "On a new class of bounds on Bayes risk in multihypothesis pattern recognition". IEEE Trans. on Computers, vol. C-23, n°1, p., Jan. 1974.

[14] **K. FUKUNAGA** : "*Introduction to Statistical Pattern Recognition*". Academic Press. 1972.

[15] **W. GERSCH** : "Nearest neighbor rule classification of stationary and nonstationary time series". in "*Applied Time Series Analysis*", t2, ed. D.F. Findley, N.Y. Academic, 1981, p. 221-270.

[16] **A.H. GRAY, J.D. MARKEL** : "Distance measures for speech processing". IEEE Trans. on Acoust., Speech, Sign. Proc., vol. ASSP-24, n°5, p. 380-391, Oct. 1976.

[17] **R.M. GRAY, A. BUZO, A.H. GRAY, Y. MATSUYAMA** : "Distorsion measures for speech processing". IEEE Trans. on Acoust., Speech, Sign. Proc., vol. ASSP-28, n°4, p. 367-376, Aug. 1980.

[18] **R.M. GRAY, D.L. NEUHOFF, P.C. SHIELDS** : "A generalization of Ornstein's $\bar{d}$ distance with applications to information theory". Annals of Probability, vol. 1, p. 315-328, 1975.

[19] **U. GRENANDER, G. SZEGÖ** : "*Toeplitz forms and their applications*". Univ. California Press, Berkeley, 1968

[20] **Y. GRENIER** : "Modélisation et reconnaissance de la parole". dans "*Outils et modèles Mathématiques pour l'Automatique, l'Analyse des Systèmes et le Traitement du Signal*". Editions du CNRS, tome 2, p. 617-637, 1982. In French.

[21] **N. ISHII, A. IWATA, N. SUZUMURA** : "Segmentation of nonstationary time-series". Int. Jal of Systems Sciences, vol. 10, n°8, p. 883-894, Aug. 1979.

[22] **N. ISHII, H. SUGIMOTO, A. IWATA, N. SUZUMURA** : "Computer classification of the EEG time-series by Kullback information measure". Int. Jal. of Systems Sciences, vol. 11, n°6, p. 677-688, June 1980.

[23] **F. ITAKURA** : "Minimum prediction residual principle applied to speech recognition". IEEE Trans. Acoust. Speech. Sign. Proc., vol. ASS-23, n°1, p. 67-72, Feb. 1975.

[24] **F. ITAKURA, S. SAITO** : "An analysis-synthesis telephony based on maximum likelihood method". Proc. Int. Cong. Acoust., c-5-5, p. c17-c20, 1968.

[25] **F. ITAKURA, T. UMEZAKI** : "Distance measure for speech recognition based on the smoothed group delay spectrum". Proc. ICASSP-87, Dallas, TX, p. 1257-1260.

[26] **R.W. JOHNSON, J.E. SHORE, J.P. BURG** : "Multisignal minimum-cross-entropy spectrum analysis with weighted initial estimates". IEEE Trans. Acoust. Speech. Sign. Proc., vol. ASSP-32, n°3, p. 531-539, June 1984.

[27] **T. KAILATH** : "The divergence and Bhattacharyya distance measures in signal selection". IEEE Trans. Communic., vol. Com-15, p. 52-60, 1967.

[28] **D. KAZAKOS** : "The Bhattacharyya distance and detection between Markov chains". IEEE Trans. on Inf. Th., vol. IT-24, n°6, p. 747-754, Nov. 1978.

[29] **D. KAZAKOS** : "Statistical discrimination using inaccurate models". IEEE Trans. Inf. Th., vol. IT-28, n°5, p. 720-728, Sept. 1982.

[30] **D. KAZAKOS, P. PAPANTONI-KAZAKOS** : "Spectral distance measures between gaussian processes". IEEE Trans. Aut. Contr., vol. AC-25, n°5, p. 950-959, Oct. 1980.

[31] **J. KITTLER** : "Mathematical methods of feature selection in pattern recognition". Int. Jal of Man-Machine Studies, vol. 7, p. 609-637, 1975.

[32] **S. KULLBACK** : "*Information theory and statistics*". Wiley, New-York, 1959.

[33] **E.L. LEHMANN** : "Testing statistical hypotheses". New-York, Wiley, 1959.

[34] **J. MAKHOUL, A.O. STEINHARDT** : "On matching correlation sequences by parametric spectral models". Proc. ICASSP-87, Dallas, TX., p. 995-998.

[35] **D. MANSOUR, B.H. JUANG** : "A family of distorsion measures based upon projection operation for robust speech recognition". Proc. ICASSP-88, New-York, NJ.

[36] **Y. MATSUYAMA** : "Mismatch robustness of linear prediction and its relationship to coding". Information and Control, vol. 47, p. 237-262, 1980.

[37] **R.K. MEHRA** : "Optimal input signals for parameter estimation in dynamic systems - Survey and new results". IEEE Trans. on Autom. Contr., vol. AC-19, n°6, p. 753-768, Dec. 1974.

[38] **N. NOCERINO, F.K. SOONG, L.R. RABINER, D.H. KLATT** : "Comparative study of several distortion measures for speech recognition". Speech Communication, vol. 4, p. 317-331, 1985.

[39] **A. V. OPPENHEIM, R.W. SCHAFER** : "*Digital Signal Processing*". Prentice Hall, 1975.

[40] **M.S. PINSKER** : "*Information and information stability of random variables and processes*". Holden Day, San Francisco, 1964.

[41] **H.V. POOR** : "Robust decision design using a distance criterion". IEEE Trans. Inf. Th., vol. IT-26, n°5, p. 575-587, Sept. 1980.

[42] **J.E. SHORE, R.W. JOHNSON** : "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy". IEEE Trans. Inf. Th., vol. IT-26, n°1, p. 26-37, Jan. 1980.

[43] **J.E. SHORE** : "Minimum cross-entropy spectral analysis". IEEE Trans. Acoust. Speech. Sign. Proc., vol. ASSP-29, n°2, p. 230-237, Apr. 1981.

[44] **J.E. SHORE, R.M. GRAY** : "Minimum cross-entropy pattern classification and cluster analysis". IEEE Patt. Anal. Mach. Intell., vol. PAMI-4, n°1, p. 11-17, Jan. 1982.

[45] **T. SODERSTROM, K. KUMAMARU** : "Some model validations criteria based on Kullback discrimination index". Proc. 24ème IEEE Conf. Decision and Control 85, p. 219-224, Fort Landerdale, Fl.

[46] **F.K. SOONG, M.M. SONDHI** : "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise". IEEE Trans. Acoust. Speech Sign. Proc., vol. ASSP-36, n°1, p. 41-48, Jan. 1988.

[47] **P.V. de SOUZA** : "Statistical tests and distance measures for LPC coefficients". IEEE Trans. Acoust. Speech Sign. Proc., vol. ASSP-25, n°6, p. 554-559, Dec. 1977.

[48] **P.V. de SOUZA, P.J. THOMSON** : "LPC distance measures and statistical tests with particular reference to the likelihood ratio". IEEE Trans. Acoust. Speech. Sign. Proc., vol. ASSP-30, n°2, p. 304-315, April 1982.

[49] **J.M. TRIBOLET, L.R. RABINER, M.M. SONDHI** : "Statistical properties of an LPC distance measure". Proc. ICASSP-79, p. 739-743.

[50] **P.M. TROUBORST, E. BACKER, D.E. BOEKEE, I.J. BOXMA** : "New families of probabilistic distance measures". Proc. 2ème Int. Joint, Conf. Pattern Recognition, Copenhague, 1974.

[51] **W. VILLEMUR, F. CASTANIE, B. GEORGEL** : "Modélisation paramétrique et classification automatique de signaux de forme transitoire". Proc. GRETSI 87, Nice. In French.

[52] **B. YEGNANARAYANA, D.R. REDDY** : "A distance measure based on the derivative of linear prediction phase spectrum". Proc. ICASSP 79, p. 744-747.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique