WAVELET OPTIMIZATION FOR CLASSIFICATION

Marie-Françoise LUCAS[1], *Christian DONCARLI* [1], *Eric HITTI* [2], *Nicolas DECHAMPS* [1]

[1]Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN)
1, rue de la Noë, BP 92101, 44321 Nantes Cedex 03, France - Marie.Lucas@irccyn.ec-nantes.fr
[2]Faculté de Pharmacie - LMPP, 2 avenue Pr. Léon Bernard 35043 Rennes cedex - Eric.Hitti@univ-rennes1.fr

ABSTRACT

This paper addresses supervised signal classification using discrete time-scale representations. Given a set of learning signals and a class of discrete wavelet-basis, we propose to select the mother wavelet which yields the best classification results. This corresponds to determining the filter (used for the decomposition) that is ideally adapted to the specific classification problem at hand. It is realized by optimizing the filter coefficients according to a contrast criterion calculated on the learning set. Simulations show the efficiency of this approach.

1. INTRODUCTION

This paper addresses supervised classification of nonstationary signals, in the realistic case where no signal model is available (as in speaker recognition, diagnosis of mechanical systems, medical diagnosis). In front of such situations, one generally maps the signals into a representation space where the discrimination between the classes is easier than in the time domain. In this paper, our objective is to optimize this mapping (or, equivalently, the transform applied to each signal) in order to obtain the best classification results, by using the information provided by the learning set. Among such approaches, we will mention a time-frequency method [4], carrying out the joint optimization of the kernel and of the distance measure by minimizing the estimated probability of classification error. Besides Saito and Coifman [9] proposed a time scale approach, aimed at computing the best wavelet basis (with respect to Fisher's criterion) from a wavelet packet decomposition. In this latter method however, the choice of the mother wavelet is not taken into account in the optimization procedure, but is fixed a priori.

We focus here on signal decomposition on a wavelet packet best basis. Our aim is to optimize the scaling filter h generating the decomposition, which is equivalent to optimize the mother wavelet. The criterion characterizing the mother wavelet quality is the estimated probability of error [4] (which is different from the empirical probability of error), computed with the signals from the learning set. In this optimization framework, a fundamental question is that of the parametrization of the filter h. In this paper, we propose some natural possibilities, however, many other possibilities can be considered.

This paper is organized as follows : after defining the problem addressed (Section 2), we present in Section 3 the scale filter parameterization used. The parameters of filter are optimized with respect to a contrast criterion. This criterion, introduced in Section 4, is the estimated probability of error. In Section 5, synthetic radar target identification signals are used to ascertain the theoretical assumptions. Moreover, we show the relevance of both the representation and the criterion selected by observing the adequacy between the contrast criterion computed from the learning set signals and the classification results obtained on a large set of test signals. Finally, we present results obtained by applying the optimization procedure and emphasize the corresponding improvement.

2. PROBLEM STATEMENT

Typically, a classification procedure requires the definition of a representation space $\mathcal{R}$, a decision rule in this space, and a quality criterion. In addition to these structural elements, a learning set composed of labelled signals is available. Let c denote the number of classes, $\mathcal{X}_i$ denote the learning set composed of the signals x belonging to the classe ω_i. Moreover, $\mathcal{Z}_i$ is the image of $\mathcal{X}_i$ in the representation space $\mathcal{R}$ (resp. z is the image of x).

In this paper, mapping x into the representation space $\mathcal{R}$ corresponds to projecting x on a wavelet packet best basis, that is, given a learning set, $\mathcal{R}$ is the space of the parameters of a given best-basis decomposition. A basis is defined by a generating filter h (called scaling filter), which is assumed to depend on a parameter vector θ (as detailed in Section 3). The space $\mathcal{R}$ is thus parameterized by θ.

The decision rule in $\mathcal{R}$ is the rule of the nearest representative, where the representative of the class ω_i (denoted $\overline{z}_i$) is computed as the average of the elements in $\mathcal{Z}_i$. According to this rule, an unknown data z will be assigned to the class ω_{i_0} that satisfies $i_0 = \arg\min_i d(z, \overline{z}_{i=1,\ldots,c})$ where d is the selected distance measure in $\mathcal{R}$.

The quality criterion (see Section 4 for a precise definition) is computed from the learning set, and the parameter θ is optimized with respect to it, which leads to minimizing the total estimated missclassification rate.

3. REPRESENTATION SPACE

In this section, we give elements on wavelet packet decomposition, then we propose a parametered filter h.

3.1. Wavelet packet best basis

A wavelet packet [7] decomposition is a generalization of the discrete wavelet transform (DWT) and of the multiresolution analysis. It is defined, as the DWT, from a scaling function $\phi(t)$ or equivalently from the scaling filter h related to ϕ by the recursive equation $\phi(t/2) = \sqrt{2}\sum_n h[n]\phi(t-n)$; the corresponding mother wavelet is defined by $\psi(t/2) = \sqrt{2}\sum_n g[n]\phi(t-n)$ with

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on December 07,2020 at 22:49:45 UTC from IEEE Xplore. Restrictions apply.

$g[k] = (-1)^{1-k}h[1-k]$ in the case of orthogonal wavelets or of a tight frame of wavelets. In the case of wavelet packet bases, the splitting of the approximation spaces is extended to the detail spaces to derive new basis, which allows adaptation to particular signals. From a scaling function $\phi(t)$, the wavelets are defined by :

$$\begin{array}{lcl} \psi_{j+1}^{2p}(t) & = & \sum_n h[n]\psi_j^p(t-2^j n) \\ \psi_{j+1}^{2p+1}(t) & = & \sum_n g[n]\psi_j^p(t-2^j n) \\ \text{with} \quad \psi_0^0(t) & = & \phi(t) \end{array} \tag{1}$$

These functions are organized according to a binary tree whose nodes (wavelet packets) are the subbasis $\Psi_j^p = \{\psi_j^p(t-2^j k)\}_{k\in Z}$ with j defining the resolution level and p the analyzed frequency band. The coefficients of the signal x decomposition on the packets are given by :

$$\begin{array}{l} c_0^0[k] = x[k] \\ c_{j+1}^{2p}[k] = \downarrow 2[c_j^p * \overline{h}][k] \\ c_{j+1}^{2p+1}[k] = \downarrow 2[c_j^p * \overline{g}][k] \\ k = 0,..,N/2^j - 1; p = 0,..,2^j - 1; j = 1,..,2^{J_{max}} \end{array} \tag{2}$$

where $\overline{f}[k] = f[-k]$, N is the length of the signal and J_{max} is the deepest level of the decomposition. The decomposition being redundant, one extracts a discriminating basis according to the Saito and Coifman algorithm [9] whose principle is the following : given a learning set $\mathcal{X}$, one calculates the discriminating capacity of each node Ψ_j^p with respect to $\mathcal{X}$; then, one seeks the set of most discriminant nodes that form a basis, according to the classical strategy of Wickerhauser and Coifman [3]. The decomposition of a discrete-time signal x, of length N, on this best basis BP^h is defined by :

$$BP_x^h = \{c_j^p[k]\}_{k=0,N/2^j-1,(j,p)|\Psi_j^p \in BP^h\}} \tag{3}$$

The exponent h indicates that the best basis depends on the generating filter. We will call the decomposition BP_x^h, possibly reduced to its most discriminating components, the *individual* z in $\mathcal{R}$.

3.2. Parameterization of the filter h

The elements of the best basis are completely determined by h and by the selection procedure carried out on the learning set according to the scheme described above. In order to generate a multiresolution analysis, h must satisfy precise conditions : for a FIR filter of length M, Lawton [5, 6] gives $M/2+1$ sufficient conditions to ensure the existence and orthogonality (or the property of being a tight frame) of the scaling function and wavelets. These conditions, which we call structural constraints, are the following :

$$\begin{array}{l} \sum_{n=1}^{M} h[n] = \sqrt{2} \\ \sum_{n=1}^{M} h^2[n] = 1 \\ \sum_{n=1}^{M} h[n]\cdot h[n-2k] = 0 \quad \text{for } k = 1,..,M/2-1 \end{array} \tag{4}$$

Therefore, there remains $M/2-1$ degrees of freedom that can be used to design the filter h for a specific application. In this paper, since the goal is classification, the optimization of the quality criterion defined in section 4 respecting the constraints (4) leads to the optimal filter h which defines a learning set-dependant mother wavelet for classification.

Moreover, regularity properties can be added to the structural constraints in order to finally obtain a smooth mother wavelet.

The most general approach consists of optimizing the M coefficients of h under the $M/2+1$ constraints given equation 4, w.r.t. to the quality criterion. However in the case of a filter of small length, it is possible to reduce the constrained problem to an unconstrained optimization over $M/2-1$ new variables. As an example [2], in the case $M = 6$, there are four structural constraints, and there remains 2 independant variables ($\theta = [a \quad b]$, $(a,b) \in [-\pi,\pi]^2$) available to define h :

$$\begin{array}{l} i = 0,1 : \\ \quad h[i] = [(1+(-1)^i\cos(a)+\sin(a)) \\ \qquad (1-(-1)^i\cos(b)-\sin(b)) \\ \qquad +(-1)^i 2\sin(b)\cos(a)](4/\sqrt{2}) \\ i = 2,3 : \\ \quad h[i] = [(1+\cos(a-b)+(-1)^i\sin(a-b))](2/\sqrt{2}) \\ i = 4,5 : \\ \quad h[i] = (1/\sqrt{2}) - h(i-4) - h(i-2) \end{array} \tag{5}$$

The results presented Section 5 are obtained with this parametric filter shape, with filters of length $M = 6$.

4. QUALITY CRITERION

In this section, we define the criterion aimed at optimizing the filter h w.r.t the learning set, then we give implementation details.

4.1. Definition of the criterion

In the context of supervised classification, a natural quality criterion is the probability of classification error, that has to be minimized. (Note that this criterion is different from the empirical error rate, which corresponds to the number of missclassified signals in the learning set, and which is proved to perform poorly.) A classification error occurs each time a signal z is assigned to ω_j while belonging to $\omega_{i\neq j}$. The probability of classification error corresponding to this case is given by (it depends on θ) :

$$P_e^\theta(j|i) = P(z \text{ assigned to } \omega_j | z \in \omega_{i\neq j})$$

and the overall probability of classification error is then :

$$P_e^\theta = \frac{1}{C}\sum_i \sum_{j\neq i} P_e^\theta(j|i)$$

Each term $P_e^\theta(j|i)$ can be calculated from the learning set as follows : let $d_{j|i}^\theta$ denote the random variable "*distance L1 of a signal from the class ω_i to the representative of the class ω_j*" and $e_{j|i}^\theta = d_{j|i}^\theta - d_{i|i}^\theta$. Then, according to the rule of nearest representative, a missclassified signal corresponds to a negative occurence of $e_{j|i}^\theta$. Under the assumption that $e_{j|i}^\theta$ follows a Gaussian distribution (this assumption is validated in Section 5), we have :

$$\begin{array}{rcl} P_e^\theta(j|i) & = & \text{Prob}(e_{j|i}^\theta < 0) \\ & = & = \frac{1}{\sqrt{(2\pi)}\sigma_{j|i}^\theta}\int_{-\infty}^{0} \exp -\frac{1}{2}(\frac{u-m_{j|i}^\theta}{\sigma_{j|i}^\theta})^2 du \\ & = & Q(\frac{m_{j|i}^\theta}{\sigma_{j|i}^\theta}) \end{array}$$

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on December 07,2020 at 22:49:45 UTC from IEEE Xplore. Restrictions apply.

with :

$$Q(u) = \frac{1}{\sqrt{(2\pi)}} \int_u^{+\infty} \exp(\frac{-t^2}{2}) dt$$
$$m^\theta_{j|i} = \mathbb{E}[e^\theta_{j|i}] \quad (\sigma^\theta_{j|i})^2 = \mathrm{Var}[e^\theta_{j|i}]$$

As usually with gaussian distributions, the quality criterion is defined by :

$$\log(\widehat{P}^\theta_e) = \log(\frac{1}{c} \sum_i \sum_{j \neq i} Q(\frac{\widehat{m}_{j|i}}{\widehat{\sigma}_{j|i}})) \qquad (6)$$

where $\widehat{m}^\theta_{j|i}$, $\widehat{\sigma}^\theta_{j|i}$ are the empirical mean and empirical variance over the learning set.

4.2. Implementation

The steps leading to the computation of the criterion are summarized below :

1. Select a value for the parameter θ (this is actually done by the optimization procedure)
2. Compute $h(\theta)$ according to definition given equation (5)
3. Decompose the signals from the learning set $\mathcal{X}$ on the wavelet packet basis induced by $h(\theta)$ according to equation (2).
4. Compute the best basis $BP^h(\theta)$; $\mathcal{Z}$ is the projection of $\mathcal{X}$ on this basis.
5. For all the signals in $\mathcal{Z}$,
   - compute $d^\theta_{j|i}$ $e^\theta_{j|i}$.
   - Estimate $\widehat{m}^\theta_{j|i}$, $\widehat{\sigma}^\theta_{j|i}$
6. Compute the criterion according to equation (6).

5. CLASSIFICATION OF CHIRPS

In this section, we show that the procedure explained above is coherent, and that both the representation and the criterion used are relevant. Moreover, we show that this new technique leads to accurate classification results by testing the same signals as in [4] (these signals correspond to the realistic application of target identification [8, 1]).

In radar target identification, when the target and the receiver are in relative motion, the radar signal can be modeled as a chirp, i.e. a quadratic phase signal whose parameters are determined by the nature of the target. We assume here that the signal observed by the receiver is two-component, namely, the signal is composed of one tone emitted by an harmonic jammer, and one chirp corresponding to the signal reflected by the target. Moreover, an additive ambient noise corrupts the observations. A sensible model is then obtained as the sum of one chirp and one tone embedded in additive white gaussian noise :

$$x[k] = A \sin 2\pi[f_0 k + \psi_0] + B \sin 2\pi[\frac{f_2 - f_1}{2N} k^2 + f_1 k + \psi_1] + b[k]$$

where $k = \{1, ..., N-1\}$, and $b[k] \sim \mathcal{N}(0, \sigma_b^2)$.

The values and probability distributions of the parameters A, f_0, ψ_0, B, f_1, f_2, ψ_1, N and SNR are for the two classes :

$\omega_1 := \{1; 0.25; \mathcal{U}[0;1]; 1; 0.4; \mathcal{U}[0.10; 0.20]; \mathcal{U}[0;1]; 128; 5dB\}$
$\omega_2 := \{1; 0.25; \mathcal{U}[0;1]; 1; 0.4; \mathcal{U}[0.25; 0.35]; \mathcal{U}[0;1]; 128; 5dB\}$

where $\mathcal{U}[\alpha, \beta]$ denotes the uniform distribution on $[\alpha, \beta]$. Note that the two classes only differ by the distribution of f_2 (this corresponds to targets having the same speed but different accelerations). Note, moreover, that the jammer and target frequencies overlap. The possible instantaneous frequencies of the chirp components are plotted on Fig. 1 for classes ω_1 and ω_2 (gray areas). We use 50 signals per class in the learning set and 1000 signals per class in the test set, both sets being disjoined.

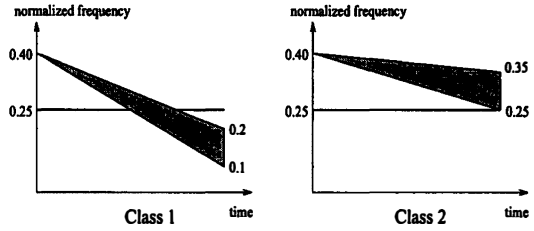

Fig. 1 – Idealized time-frequency representations of the two classes.

We firstly show that the assumption of Gaussian $e^\theta_{j|i}$ is correct. Given a representation space, i.e. given a parameter θ, Fig. 2 displays the distribution of the random variable $e^\theta_{1|2} = d^\theta_{1|2} - d^\theta_{2|2}$ for 1000 signals : this clearly demonstrated that the Gaussian assumption is realistic, and, as a consequence, that the criterion definition given in Eq. (6) is relevant.

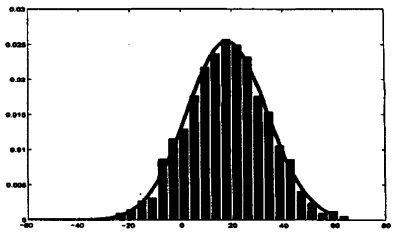

Fig. 2 – Distribution of $e^2_{1|2}$ estimated on a set of 1000 signals.

Secondly, we show that the criterion used for the optimization is consistent, i.e., that it indicates the actual performance of the representation. The proposed procedure is relevant only if minimizing the criterion corresponds to selecting a representation space with a good capacity of generalization, that is, $\mathcal{R}$ performs well in classifying the test signals. An illustration of these performances is given Fig. 3 : for each point, the first coordinate is the missclassification rate obtained with a test set and the second coordinate is the value of $\log(\widehat{P}^\theta_e)$ computed by using the learning set. The points are generated, on one hand, (points represented by a dot '.') by taking different filters h of length $M = 6$ obtained by a random selection of the parameter $\theta = [a, b]$ (see Eq. (5)) and , on the other hand, (points represented by a cross '+') by using predetermined filters of various lengths and classical generators of wavelets (Daubechies, splines, coiflets).

Thirdly, we present classification results obtained with an optimal filter. Fig. 4 displays the value of $\log(\widehat{P}^\theta_e)$ as a function of

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on December 07,2020 at 22:49:45 UTC from IEEE Xplore. Restrictions apply.

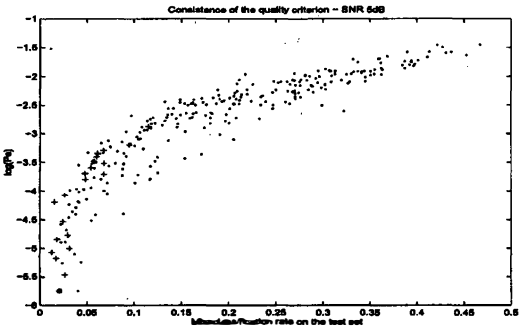

Fig. 3 – Generalization capacity of both the criterion and the representation. Globally, minimizing $\log(\widehat{P}_e^{\theta})$ computed from the learning set leads to minimizing the missclassification rate computed from the test set. Each point (.) corresponds to a randomly selected filter h of length 6 and each cross (+) corresponds to generating filters of length $M > 6$, leading to a classical wavelet. The star (*) shows the performances of the optimal 6-length filter $h(\theta_{opt})$.

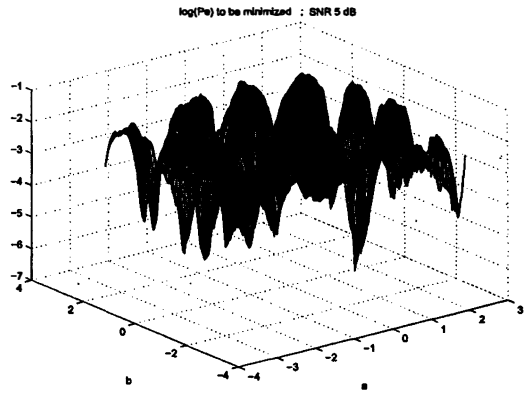

Fig. 4 – $\log(\widehat{P}_e^{\theta})$ versus the parameters (a, b), a and b being discretized with a step of 0.01 over $[-\pi, \pi]$.

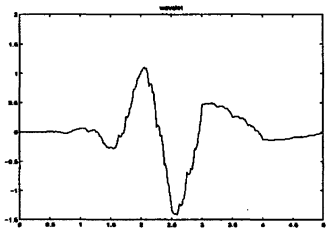

Fig. 5 – Mother wavelet corresponding to the optimal filter $h(\theta_{opt})$. This approximation is calculated from h by using the iterative algorithm described in [2].

discretized $\theta = (a, b)$, the parameters to be optmized. This function exhibits a lot of local minima, which indicates that a global optimization procedure must be used. Moreover, several local minima correspond to very close values of the criterion (this is due to the small size of the learning set).

So we prefer to select a minimum leading to a smooth wavelet, by adding a constraint on the first moment of the filter g (the wavelet filter, defined in Section 3.1).

Consequently, a constrained simulated annealing procedure has been implemented to optimize θ under this regularity constraint. The starting point, $\theta_{init} = [0, 0]$ (no constrained) corresponds to the Haar wavelet. The value of the criterion $\log(\widehat{P}_e^{\theta})$ in this initial point is -2.3 and the missclassification rate computed from the test set is 23%. For the selected optimum ($\theta_{opt} = [-1.3416, 1.0584]$), represented by a star on Fig. 5, the value of the criterion is -5.8 and the missclassification is now 2%, which is good compared to the results given in [4] (where the same signals are processed). The corresponding optimal mother wavelet is represented Fig. 5.

6. CONCLUSION

We have shown that optimizing the shape of the mother wavelet with respect to the parameter θ improves significantly the classification performance. Moreover, it is a substantial additional contribution to a simple search for best basis. In addition, Fig. 3 shows that, even if the optimal solution is restricted to the space of filters of length 6, this solution behaves very often better (in terms of capacity of generalization) than those corresponding to classical wavelets defined from filters of length greater than 6, and having usually good regularity properties. This corresponds to excellent representation capacities. Extensions to more general generative filters are under study.

7. REFERENCES

[1] O. Besson, M. Ghogho, A. Swami. *Parameter estimation for random amplitude chirp signals*. IEEE trans. Signal Processing, vol. 47, no. 12, pp. 3208-3219, Dec. 1999.

[2] C.S. Burrus, R.A. Gopinath, H. Guo. *Introduction to wavelets and wavelet transforms*. Prentice Hall, 1998.

[3] R.R. Coifman, M.V. Wickerhauser. *Entropy based algorithm for best basis selection*. IEEE Transaction on information theory, 38 :713-718, 1992.

[4] M. Davy, C. Doncarli, F. Boudreaux. *Improved Optimisation of Time-Frequency Based Classifiers*. IEEE Signal Processing letters, Vol 8 n°2, pp. 52-57, 2001.

[5] W. Lawton. *Tight frames of compactly supported affine wavelets*. Journal of Mathematical Physics, Vol 31 n°8, pp. 1898-1901, 1990.

[6] W. Lawton. *Necessary and sufficient conditions for constructing orthonormal wavelet bases*. Journal of Mathematical Physics, Vol 32 n°6, pp. 1440-1443, 1991.

[7] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

[8] A. Rihaczec. *Principles of high-resolution radar*. McGraw-Hill, New York, 1969

[9] N. Saito, R.R. Coifman. *Local discriminant bases*. Wavelet applications in Signal and Image Processing II,A.F. Laine and M.A. Unser, Eds. Proc. SPIE vol 2303, 1994.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on December 07,2020 at 22:49:45 UTC from IEEE Xplore. Restrictions apply.