A local feature extraction using biorthogonal bases for classification of embedded classes of signals

P. Terwiesch[†] P. Mercorelli[‡]
[†]Automation Technologies – ABB Corporate Research
Segelhof, CH-5405 Baden (Switzerland)
Tel. +41 56 486 8319 – Fax +41 56 486 7365
Peter.Terwiesch@ch.abb.com
[‡]Automation and Information Technology Department – ABB Corporate Research.
Speyerer Str. 4, D-69115 Heidelberg (Germany).
Tel. +49 6221 596263 – Fax +49 6221 596253
Paolo.Mercorelli@de.abb.com
[‡]Dip. di Ingegneria dell' Informazione – Università di Siena
via Roma, 56, 53100 Siena, Italia.
mercorelli@ing.unisi.it

Keywords: signal classification, trigonometric bases, wavelet packets, wavelet networks, nonparametric regression.

1 Abstract

We describe a method to construct orthonormal basis coordinates which maximizes over redundant dictionaries (frames) of biorthogonal bases the class separability index or *distances* among classes. The method proposes an algorithm which consists of biorthogonal expansions over two *redundant dictionaries*. In multiclassification problems, *embedded classes* are often present, we show how the biorthogonality of the expansion can really help to construct a coordinate system which characterizes the classes. The algorithm is created for training wavelet networks in order to provide an efficient coordinate system maximizing the *Cross Entropy function* between two complementary classes. The algorithm works with a preliminary extracted features with shrinkage technique in order to reduce the dimensionality of the problem. In particular, our attention is pointed out for a practical time frequency monitoring, detection and classification of transients in rail vehicle systems. We want to distinguish transients as *inrush current* and *no inrush current* and among them between the two complementary classes: *dangerous inrush current* and *no dangerous inrush current*. The proposed algorithm could be used on line in order to recognize the dangerous transients in real time and thus shut-down the vehicle. We show how, with a limited number of wavelets and with few iterations on the compressed data, good and fast performances are achieved. Simulations using real measured data on the vehicle line are included to illustrate and to analyse the effectiveness of the proposed method.

2 Introduction and Problem Definition

Extracting relevant features from signals or images is an important process for data analysis. In particular when we are interested to classifying signals into known categories. We propose a method for detection and classification of time-frequency phenomena by using wavelet packets in wavelet networks with biorthogonal activation functions. Specifically we are interested to classify transient harmonics as they occur in electrical power systems [1, 3]. This literature has indicated wavelets and wavelet networks as a promising approach for off line analysis and classification. More, a recent work in this direction [2] developed an efficient algorithm in order to detect and classify transient harmonic phenomena as *inrush current* which presents an unimodal distribution. This paper tries to extend previous results as in [2], by several structural modifications on the training algorithm, to the class of *embedded phenomena* with which, very often, can be modeled a classification system. In other words we can think about embedded phenomena such as the following structured classes.

Definition 1 *Given a set $\mathcal{Y}$ of $(L+1)$ classes of signals: $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_L, \mathcal{C}_C\}$, where L are known classes and $\mathcal{C}_C$ is the class defined as complement to the class $\mathcal{C}_L$. We will call the set $\mathcal{Y}$ a set of* embedded classes *if the following relationships hold:*

$$\mathcal{C}_L \cap \mathcal{C}_C = \{0\} \tag{1}$$

and

$$\begin{array}{c} \mathcal{C}_1 \cap \mathcal{C}_2 = \{0\}; \quad \mathcal{C}_1 \cup \mathcal{C}_2 = \mathcal{C}_3; \\ \mathcal{C}_3 \cap \mathcal{C}_4 = \{0\}; \quad \mathcal{C}_3 \cup \mathcal{C}_4 = \mathcal{C}_5; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \mathcal{C}_{(L-2)} \cap \mathcal{C}_{(L-1)} = \{0\}; \quad \mathcal{C}_{(L-2)} \cup \mathcal{C}_{(L-1)} = \mathcal{C}_L. \end{array} \tag{2}$$

In our specific task the embedded classes could be seen in the following way:

Let L=3, and let $\mathcal{C}_1 = \{No\, dangerous\, inrush\, current\}$ and we will pose its relative complementary class such as $\mathcal{C}_2 = \{Dangerous\, inrush\, current\}$ and thus at the end $\mathcal{C}_3 = \{Inrush\, current\}$ with its relative complementary class $\mathcal{C}_C = \{No\, inrush\, current\}$. Obviously one has that $\mathcal{C}_3 \cap \mathcal{C}_C = \{0\}$ and $\mathcal{C}_1 \cup \mathcal{C}_2 = \mathcal{C}_3$. We want to classify the inrush current which for instance has an energy which exceeds several limits in several particular bands of frequency. In [2] we showed as, looking for the best basis (Shannon Entropy function) over a dual frame (smooth trigonometric sine/cosine wavelet functions), a dual coordinate system can be effectively determined using a particular training algorithm for wavelet networks. In [2] the task was: *detect and classify inrush current*.

In this paper the problem looks a little bit more complicated. These kind of phenomena show several difficulties to be classified because of their physical characteristics. In this sense it is very important the a priory knowledge on the phenomenon in order to choose an appropriated family regressor for extracting relevant features. This reduces the dimensionality of the problem without loosing information and constructs at the mean time an efficient and discriminat system of coordinate. We want to focus our attention on one method of selection of coordinate systems through training wavelet networks. We proposed an algorithm which works with biorthogonal wavelet functions and in particular we posed our attention over a harmonic detection problem.

Now, why we are using wavelet packets and their relative biorthogonal families in nonorthogonal libraries ?

We need to define a *language* to describe the signals. More, we have to seek of optimizing certain criteria depending on the particular problem. The language must be as much versatile as possible and as much elastic as possible in order to describe various local physical features of the *embedded classes of signals*. In the mean time the method must be computationally efficient to be practically applied. The wavelet networks provide flexible coordinate systems, their adaptive time frequency cells are able to capture the features of the signals in a reasonable computational calculations. More, the nonorthogonal libraries allow more elasticity in order to approximate the training signals, see [4], even though the optimal expansion of a signal in a redundant dictionary of waveforms is an NP-complete problem. At the end we want to remark that the biorthogonaliy of the frames plays a particular role in order to perform an orthogonal coordinate system using the well known *Cross Entropy function* which is a measure of the discrepancy between two or more bases. The algorithm, working on a preliminary compression data (*data shrinkage*), builds a basis which *illuminates* the differences among embedded classes. It describes every couple of complementary classes with biorthogonal bases in order to perform an orthonormal coordinate system maximizing, during every step, the *Cross Entropy function*, see section 4. In this way complementary phenomena are considered such as bimodal phenomena and the biorthogonal approximating bases allow to perform for each complementary class two orthonormal coordinate basis systems. Wavelet networks seem to provide a natural way to attack our problem even though several basic issues schould be well satisfied.

2.1 Problem Statement

Let $\mathcal{X} \subseteq \Re^{d_0}$ where d_0 is the number of samples for each signal and the dimensionality of the *signal space* ($\mathcal{X}$) or pattern space which is a subset of the standard d-dimensional vector space which contains all signals under consideration. In this case, the dimensionality of the signal space is equivalent to the length of the signals, we are working with dyadic wavelet packets and thus we will assume that $d_0 = 2^{n_o}$ for some n_o. As in (1), let $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_L, \mathcal{C}_C\}$ be a set of L known categories or classes and $\mathcal{C}_C$ defined as complement, in general for regression problems $\mathcal{Y} = \Re$. Let N_l be the number of signals belonging to the class l, i.e. let us denote a set of class l signals by $\{\mathbf{x}^l\}_{i=1}^{N_l}$. Now the signal classification problem can be considered as:

find a map called feature extractor $f : \mathcal{X} \rightarrow \mathcal{F} \subset \Re^k$ *(normally more than one) with* $k \ll d_0$ *in order to extract relevant features.*

This reduces the dimensionality of the problem under consideration without loosing important information but improving the efficiency of the pattern recognition.

It is worthwhile to pay attention to the dimensionality problem, in order to improve the efficiency of the algorithm. Nonparametric estimators soffer from the curse of dimensionality, the complexity of the estimators grows with the input dimensional d_0. The phenomenon is related to the fact that the sample length is exponential related to the length d_0 of the input sample $\mathcal{X}$. In general when the input data training are sparse in the input space $\Re^{d_0}$ and only part of this space 'explored' by the sparce training data are of interest then the curse of dimensionality is not so heavy. This normally happens in classification problems and modeling of control system where the estimation is done only on the same small portion of the input space. In order to overcome the curse of dimensionality problem it is also useful a good choice of the estimator, when the estimator is *close* to the considered physical phenomenon then the *information is concentrated in few data*. In our case the reduction of the dimensionality is performed in two steps.

Step 1: *data shrinkage* in order to reduce the number N_l of the signals (the sample length).

Step 2: compression of the prototype signals resulting from the training technique with the wavelet network in order to have the time frequency cell basis coordinates.

The step 1 allows us to select the most important subspaces in fact, given an approximating wavelet library not all the wavelet functions are useful but only a small number of these are important. In particular there are several sets of signals which don't hit the wavelt support and these can be eliminated. The step 2 compress the information into the best time-frequency probability distribution subspaces.

Once the map is built the classification problem could consist of evaluating the Euclidean distance between the projected fresh signal through the algorithm and the feature coordinate basis vectors (compressed prototypes). In general the recognition problem can be written as a map (generally nonlinear) $g : \mathcal{F} \rightarrow \mathcal{Y}$. The proposed algorithm uses basically the best basis paradigm as in [8] or [11] which allows a rapid search among a large collection of bases (the complexity of the calculation is $\mathcal{O}(k\big(log(k)\big)^p)$, where p is equal 1 or 2 depending on the basis type, wavelet dictionaries or trigonometric wavelet dictionaries respectively).

The paper is organized as follows. In section 3 we discuss briefly the smooth trigonometric wavelet packets and several problems connected to the choice of the best regressor family. In section 4 we bresent the algorithm with its mathematical details. In section 5 we report the achieved results.

3 Background and Several Issues

Wavelet transform and wavelet series are becoming popular in signal processing and numerical analysis. Loosely speaking, a function $f(t)$ can be decomposed into

$$f(t) = \sum_j \sum_n w_{j,n} \psi_{j,n}(t) \quad (3)$$

where the $\psi_{j,n}(t)$ are the wavelet functions, normally obtained by dilating and translating a mother function $\psi(t)$, the index j and n denote the dilation and translation respectively and $w_{j,n}$ is the weight coefficient for $\psi_{j,n}(t)$. The most popular algorithms are related to the orthonormal wavelet bases, see [7], characterized from fast and elegant algorithms. There are, besides these, less used, the *wavelet frames*, see [7], for which the computations of the coefficients are more complicated but which have certain advantages. As wavelet frames consist of nonorthogonal wavelet families, they are *redundant bases*. To be more formal:

Definition 2 *A family of functions* $\{\psi_{j,n}(t); (j,n) \in \mathsf{Z}, t \in \Re\}$ *in a Hilbert space* $\mathcal{H}$ *space is called a frame of* $\mathcal{H}$ *if for every element* $f(t) \in \mathcal{H}$ *there are two positive constants* $\mathbf{A}$ *and* $\mathbf{B}$ *such that:*

$$\mathbf{A}||f(t)||^2 \leq \sum_{j,n} ||\langle f(t), \psi_{j,n}(t)\rangle||^2 \leq \mathbf{B}||f(t)||^2. \quad (4)$$

Where with $\langle \cdot, \cdot \rangle$ we have indicated the inner product and with $|| \cdot ||$ the norm.

In this approach we have a drawback, the optimal decomposition on a nonorthogonal basis is a NP-complete problem and we need to stop the algorithm, for instance, with a threshold criterion at the stage i for the $\mathcal{L}^2$ norm of the differential error, we will come back later to the question.

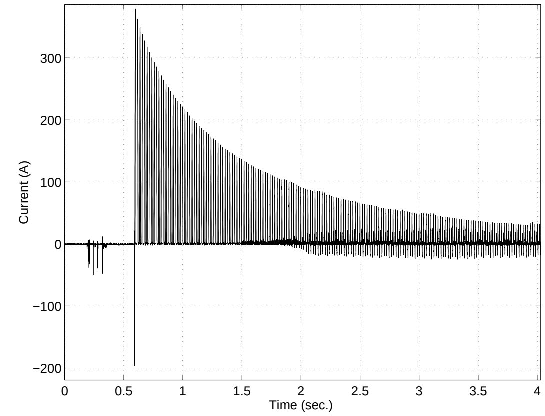

Figure 1: Real Signal. Inrush current: time domain. Right: Inrush current: windowed spectral analysis.

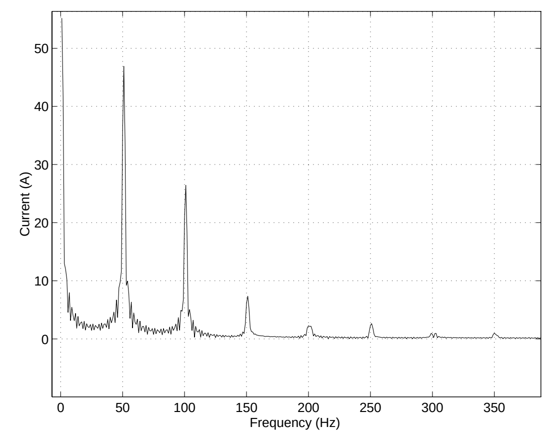

Figure 2: Real Signal. Inrush current: windowed spectral analysis.

Now, the first question is which function to use like activation function. This is a difficult decision, the collected experience on this sense doesn't help too much. All of the model structures are capable of approximating any 'reasonable function' [7]. Thus the question is pick one that 'suits the application', in the sense that only few terms will be needed. A suitable criterion known in the literature is to select the basis which, once fixed a threshold level, has the minimum number of elements in the selected frame. Now, having chosen the best family how to choose the size of the frame subset ? Finally, how we can select the terms of the subset ?

3.1 Choosing the Best Family Regressor

The case presented in this paper has quasi-harmonic signals that change amplitude and phase over time. This latter aspect suggests the wavelet like activation function. We show in Fig. 1 and Fig. 2, where are depicted a measured signal in time domain and its windowed Fourier transform respec-

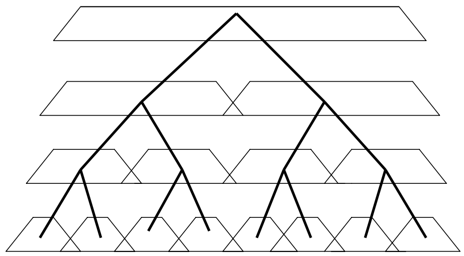

Figure 3: Organization of local intervals into a binary tree for smooth local trigonometric wavelet. Orthogonal basis.

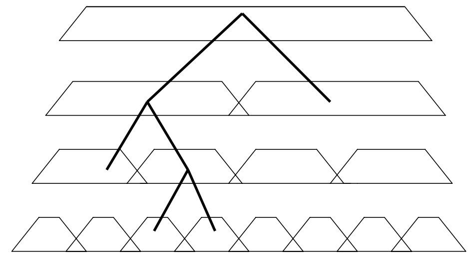

Figure 4: Organization of local intervals into a binary tree for smooth local trigonometric wavelet. Frame.

tively, how the data are very well concentrated around several frequencies, in this case they are multiple of the fundamental (50 Hz). The picture in Fig. 2 seems to suggest a function with a frequency window and time support. As shown in [2] a suitable family for this case is the smooth trigonometric wavelet packet. We want just to recall several basic aspects, further details in [10].

Definition 3 *Let a library of wavelet packets be the collection of functions of the form*

$$\psi_{(d,j,n)}(t) = \psi_j(2^d t - n) \tag{5}$$

where $(d,n) \in \mathsf{Z}$ *and* $j \in \mathsf{N}$.

We have already remarked that we are talking about truncated indices, thus finite libraries of wavelet packets. Here, the *pyramidal* packet is represented with the indices (d,j,n), d is the level of the tree (scaling parameter), j is the frequency cell (oscillation parameter) and n the time cell (localization parameter).

We consider a cover of the real axis $\Re = \bigcup_{-\infty}^{\infty} \mathcal{I}_i$, where $\mathcal{I}_i = [\alpha_i, \alpha_{i+1})$ and $\alpha_i < \alpha_{i+1}$.

Write $\mathcal{T}_i = \alpha_{i+1} - \alpha_i = |\mathcal{I}_i|$ and let $\mathcal{W}_i(t)$ be a window function supported in $[\alpha_i - \frac{\mathcal{T}_{i-1}}{2}, \alpha_{i+1} + \frac{\mathcal{T}_{i+1}}{2}]$ such that

$$\sum_{-\infty}^{\infty} \mathcal{W}_i^2(t) = 1 \tag{6}$$

and

$$\mathcal{W}_i^2(t) = 1 - \mathcal{W}_i^2(2\alpha_{i+1} - t) \qquad for \quad t \quad near \quad \alpha_{i+1}. \tag{7}$$

These conditions tell how the bell function should be taken in order to ensure the orthogonality of the basis, see for instance *Lemma 3* in [6]. This shows that choosing a basis as in Fig. 3, adiacent function, we abtain a orthonormal basis. On the contrary if we consider bases on different levels of the tree as in Fig. 4 these don't form an orthonormal basis

The functions

$$\mathbf{S}_{i,k}(t) = \frac{2}{\sqrt{2\mathcal{T}_i}} \mathcal{W}_i(t) sin[(2k+1)\frac{\pi}{2\mathcal{T}_i}(t - \alpha_i)] \tag{8}$$

form an orthonormal basis of $\mathcal{L}^2(\Re)$ subordinate to the partition $\mathcal{W}_i$. The collection of such bases forms a library of orthonormal bases [10]. We can form a library of orthonormal local cosine bases:

$$\mathbf{C}_{i,k}(t) = \frac{2}{\sqrt{2\mathcal{T}_i}} \mathcal{W}_i(t) cos[(2k+1)\frac{\pi}{2\mathcal{T}_i}(t - \alpha_i)]. \tag{9}$$

We have to remark that taking equal smooth windows $\mathcal{W}_i(t)$ (see [10]) then sine/cosine orthogonality can be maintained, see Fig. 5 and Fig. 6 where we have depicted the sine/cosine bases related to the second level of the packet tree with a frequency of 50 Hz.

It is easy to check that if $\mathbf{H}_i$ denotes the space of functions spanned by $\mathbf{S}_{i,k}$ for $k = 0, 1, 2, ...$ then $\mathbf{H}_i + \mathbf{H}_{i+1}$ is spanned by

$$\mathbf{S}_{i,k}(t) = \mathcal{P}(t) sin[(2k+1)\frac{\pi}{2(\mathcal{T}_i + \mathcal{T}_{i+1})}(t - \alpha_i)], \tag{10}$$

where

$$\mathcal{P}^2(t) = \frac{1}{\sqrt{2(\mathcal{T}_i + \mathcal{T}_{i+1})}} \left(\mathcal{W}_i^2(t) + \mathcal{W}_{i+1}^2(t)\right)$$

is a 'window' function covering the interval $I_i \cup I_{i+1}$.

It is necessary to see the connections between the localized trigonometric functions and wavelet packets. If we consider the frequency line $\Re$ split as the union of $\Re^+ = [0, +\infty)$ and $\Re^- = (-\infty, 0)$. On $\mathcal{L}^2(\Re^+)$ we introduce a window function $\mathcal{W}(\omega)$ such that

$$\sum_{k=-\infty}^{\infty} \mathcal{W}^2(2^{-k}\omega) = 1. \tag{11}$$

Clearly we can view $\mathcal{W}(2^{-k}\omega)$ as window function above the interval $(2^k, 2^{(k+1)})$ and observe that the functions

$$\mathbf{s}_{k,j}(\omega) = \mathcal{W}(2^{-k}\omega) sin[\Big(j + \frac{1}{2}\Big)\pi\Big(\frac{\omega - 2^k}{2^k}\Big)] \tag{12}$$

form an orthogonal basis of $\mathcal{L}^2(\Re^+)$. Similarly

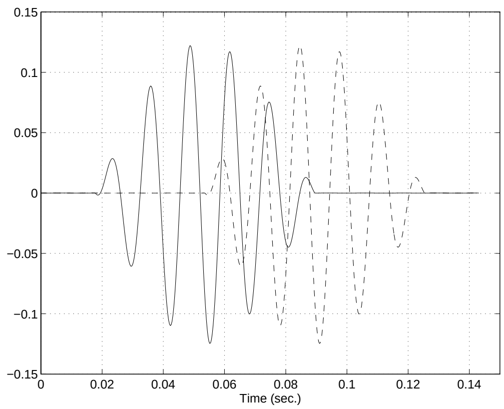

Figure 5: Adjacent (orthogonal) cosine waveforms with smooth window $\mathbf{C}_{(2,1)}(t)$ and $\mathbf{C}_{(2,2)}(t)$.

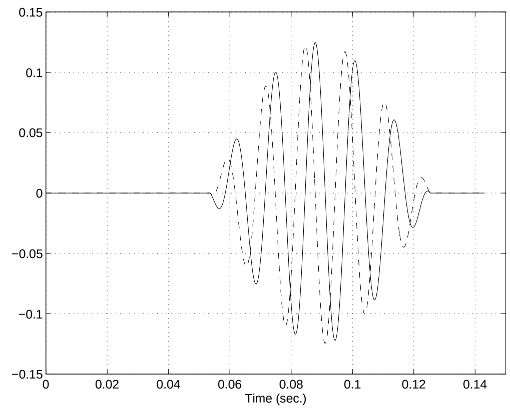

Figure 6: Biorthogonal smooth local sine and cosine function $\mathbf{S}_{(2,2)}(t)$ and $\mathbf{C}_{(2,2)}(t)$.

$$\mathbf{c}_{k,j}(\omega) = \mathcal{W}(2^{-k}\omega)cos[\Big(j+\frac{1}{2}\Big)\pi\Big(\frac{\omega-2^k}{2^k}\Big)] \qquad (13)$$

gives another orthonormal basis but whose elements are not orthogonal to the functions $\mathbf{s}_{k,j}$.

In order to build the biorthogonal $\mathbf{S}_{k,j}$ sine and $\mathbf{C}_{k,j}$ cosine bases we define $\mathbf{S}_{k,j}$ as an odd extension to $\Re$ of $\mathbf{s}_{k,j}$ and $\mathbf{C}_{k,j}$ as an even extension of $\mathbf{c}_{k,j}$. In this way we can find $\mathbf{S}_{k,j} \perp \mathbf{C}_{k^{'},j^{'}}$ for every $k, j, k^{'}, j^{'}$, permitting us to write:

$$\mathbf{C}_{k,j} \pm i\mathbf{S}_{k,j} = \exp\big(\frac{\pm ij\pi\omega}{2^k}\big)\tilde{\psi}(\omega)$$

where $\tilde{\psi}(\omega)$ is the Fourier transform of the Meyer wavelet $\mathbf{\Psi}$ defined in [12], further details of this calculation may be found in [10]. In Fig. 5 and in Fig. 6 we have depicted the sine and cosine biorthogonal functions in the time domain, where we have built the functions with the prototype cutoff $\mathcal{W}(t)$ like a sine function available in [5].

It easy to see how the time and frequency cells are linked in a dyadic way, this sort of analysis is equivalent to wavelet packet analysis which allows us to perform an adapted Fourier windowing directly in the time domain. The wavelet packet library is constructed by iterating the wavelet algorithm. It easy to see how the time and frequency cells are linked in a dyadic way, this sort of analysis is equivalent to wavelet packet analysis which allows us to perform an adapted Fourier windowing directly in the time domain. The wavelet packet library is constructed by iterating the wavelet algorithm.

From now on and in order to formally define the *library of wavelet packets* we will consider a new index notation.

Definition 4 *Let a library of wavelet packets be the collection of functions of the form*

$$\psi_{(d,j,n)}(t) = \psi_j(2^d t - n) \qquad (14)$$

where $(d,n) \in \mathsf{Z}$ *and* $j \in \mathsf{N}$.

We have already remarked that we are talking about truncated indices, thus finite libraries of wavelet packets. Here, the *pyramidal* packet is represented with the indices (d, j, n), d is the level of the tree (scaling parameter), j is the frequency cell (oscillation parameter) and n the time cell (localization parameter).

The function $\psi_{(d,j,n)}(t) = \psi_j(2^d t - n)$ is roughly centered at $2^{-d}n$, has support of size $\approx 2^{-d}$ and oscillates $\approx j$.

To go a little bit more in depth, we suppose that the signal consists of $N = 2^{N_0}$ dyadic and equally spaced samples and the library tree contains all the local trigonometric analyses to level N_0 of the frame, with windows of size $2^{N_0}, 2^{N_0-1}$, ...1. The basis function will be indexed by the triplet (d, j, n): if N is the total number of the samples then the corresponding samples related to the d *level* with relative desampling are $N_d = 2^d$ and $0 \leq d \leq N_0$, $0 \leq j < 2^{N_0-d}$, $0 \leq n < 2^d$.

The scale parameter d gives the number of decompositions of the original signal window into subwindows and the position index n numbers the adjacent windows. *Thus the information cell is drawn over the horizontal (time) interval* $I_n = [2^{N_0-d}n, 2^{N_0-d}(n+1)[$. In general, the local trigonometric bases, for instance the cosine basis, for the subspace over the time subinterval I_n consists of the function with the associated information cell alongside the frequency interval $I_j = [2^d j, 2^d(j+1)[$ on the vertical axis (frequency). The basis functions have the nominal frequencies in $2^d(j+\frac{1}{2})$. Each subdivision halves the nominal window width and thus the resolution level, in particular the resolution level on the tree could be represented like a collection of rectangles:

$$[2^{N_0-d}n, 2^{N_0-d}(n+1)[\times[2^d j, 2^d(j+1)[. \qquad (15)$$

Taking a basis with cells on different level of the tree we obtain a nonorthogonal basis (frames): the symmetry of the windows is lost but not their derivability, they are sums of the derivable functions. In the other words, taking basis elements on different levels of the tree which cover the real axis $\Re$ we are considering superpositions of bases with different resolution frequency cells, the orthogonality is lost. Our algorithm will work transversally on the wavelet packet tree without any restriction in order to use all the possible combinations of the bases, all the possible frames. Once selected the family regressor, for instance the truncated sine/cosine wavelets, the (d, j, n) parameterized family:

$$\left\{\mathcal{R}_c, \mathcal{R}_s\right\} = \left\{\psi^c_{(d,j,n)}(t), \psi^s_{(d,j,n)}(t); (d,n) \in \mathsf{Z}, j \in \mathsf{N}, t \in \Re\right\}$$

should contain a finite number of wavelets, as less as possible, so that the regressor selection procedure can be efficiently applied. Given an approximating wavelet library not all the wavelet functions are useful, normally only a small number of the coefficients are important, the other ones can be neglected. In order to explain the construction of wavelet network let us start with a regular wavelet lattice. Many wavelets in the regular lattice do not contain any data point in their support because of the sparseness of the data. The training data point don't provide any information for determining the coefficients of these empty wavelets, this means that they are superfluous for the regression estimation and could be eliminated. In general we can select the candidate library as follows:

$$\left\{\mathcal{R}_c,\ \mathcal{R}_s\right\} = \left\{\psi^c_{(d,j,n)}(t), \psi^s_{(d,j,n)}(t) : (d,j,n) \in \mathcal{I}_1 \cup \mathcal{I}_2 \cup ....\cup \mathcal{I}_K\right\} \tag{16}$$

with $K = 1, 2.......L$ and

$$\mathcal{I}_k = \left\{(d,j,n) : \|\psi^c_{(d,j,n)}\|_p > \epsilon, \|\psi^s_{(d,j,n)}\|_p > \epsilon\right\}, \tag{17}$$

where ϵ is a chosen small positive number. In this way the 'empty' wavelets are eliminated from the wavelet frame. In other words we are starting from a regular tree packet (library) and we select only those which their support hit our training data. This method is called by some authors *wavelet shrinkage* [9]. We will show that with very few bases of the local trigonometric functions we can obtain a good function detector.

4 The Proposed Algorithm

The extractor matrix $\mathbf{A} : \mathcal{X} \to \mathcal{F} \subset \Re^k$ ($\mathbf{A} : \Re^{d_o} \to \Re^k$) is an unitary matrix[1] see [2], which is a map of the wavelet packet trees. Now is known that, given a subspace $\mathcal{H} \subset \mathcal{F}$ of $\Re^k$ and let H be an unitary basis matrix of $\mathcal{H}$, then the orthogonal projection matrices on $\mathcal{H}$ and on $\mathcal{H}^\perp$ are $\mathbf{P} = \mathbf{H}\mathbf{H}^T$ and $\mathbf{Q} = \mathbf{I} - \mathbf{H}\mathbf{H}^T$, respectivaly. It easy to see that, given an $x \in \mathcal{H}$ then $\mathbf{P}x = x$ and given $y \in \mathcal{H}^\perp$ then $\mathbf{P}y = 0$. These properties suggest to perform two orthogonal subspaces which characterize every two complementary classes choosing among all the possible bases which maximize the Cross Entropy function. The Cross Entropy function is a measure of the discrepancy among L distributions, normally one may take $\binom{N_0}{2}$ pairwise combinations of this measure. If we call $\mathbf{p}^{(1)}, ..., \mathbf{p}^{(L)}$ the L distributions and with $\mathcal{D}$ the discrepancy measure then:

$$\mathcal{D}(\{\mathbf{p}_l\}_l^L) = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \mathcal{D}(\mathbf{p}^i, \mathbf{p}^j).$$

The *additive* function $\mathcal{D}$ for two distribution $\mathbf{p}$ and $\mathbf{q}$ is defined as

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) = \sum_k^{N_{l_1}} \sum_h^{N_{l_2}} \mathbf{p}_k \ln\left(\frac{\mathbf{p}_k}{\mathbf{q}_h}\right),$$

where N_{l_1} and N_{l_2} are the number of the element for each class and the indices k and h are the internal indices, further details in [8].

4.1 Mathematical Details

In order to consider and to use the nonorthogonality of the frames which generates an interaction between the elements of the bases[2] the algorithm considers to every step all the elements of the bases previously selected, without any elimination, see [4]. Because of the decomposition on a nonorthogonal basis is not unique we need to stop the algorithm, for instance, with a threshold criterion at the stage i for the $\mathcal{L}^2$ norm of the differential error.

The algorithm can be mathematically represented as follows.
Let

$$\left\{\mathcal{R}_c, \mathcal{R}_s\right\} = \left\{\psi^c_{(d,j,n)}(t), \psi^s_{(d,j,n)}(t); (d,n) \in \mathsf{Z}, j \in \mathsf{N}, t \in \Re\right\}$$

be the truncated cosine and sine packet frames respectively as defined in (16).

0. Define the initial residual $\gamma_{c(0)}(k) = \mathbf{x}_{l_{1k}}$, $k = 1, 2, ..., N_{l_1}$ and $\gamma_{s(0)}(h) = \mathbf{x}_{l_{2h}}$, $h = 1, 2, ..., N_{l_2}$. Where the $\mathbf{x}_{l_{1k}}$ and $\mathbf{x}_{l_{2h}}$ are the observed signals as defined in section 2. Fixed a *stage index* M such that minimizes the residual $\mathcal{L}^2$ norm as above mentioned, let $\mathbf{f}_{c_0}(t) = 0$ and $\mathbf{f}_{s_0}(t) = 0$.

[1] An $\mathbf{A}$ unitary matrix is such that $\mathbf{A}^{-1} = \mathbf{A}^T$.

[2] In a frame the decomposition is not unique.

Begin-loop

1. For $i = 1, 2, ..M$.

Calculate the weights $\mathbf{c}_{(d,j,n)}(k)$, $\mathbf{s}_{(d,j,n)}(h)$ on all cosine and sine wavelet packet trees according to the index:

$$\mathcal{J}\Big(\mathbf{c}_{(d,j,n)}(k), \mathbf{s}_{(d,j,n)}(h)\Big) =$$

$$\frac{1}{N_{l_1}}\sum_{k=1}^{N_{l_1}}\Big(\gamma_{c(i-1)}(k) - \sum_{(d,j,n)\in\mathcal{R}_c}\mathbf{c}_{(d,j,n)}(k)\psi^c_{(d,j,n)}(k)\Big)^2 +$$

$$\frac{1}{N_{l_2}}\sum_{h=1}^{N_{l_2}}\Big(\gamma_{s(i-1)}(h) - \sum_{(d,j,n)\in\mathcal{R}_s}\mathbf{s}_{(d,j,n)}(h)\psi^s_{(d,j,n)}(h)\Big)^2,$$

this yields:

$$\mathbf{c}_{(d,j,n)} = \Big(\frac{\sum_{(d,j,n)\in\mathcal{R}_c}\langle\gamma_{c(i-1)}(k), \psi^c_{(d,j,n)}(k)\rangle}{\Big(\sum_{(d,j,n)\in\mathcal{R}_c}\big(\psi^c_{(d,j,n)}(k)\big)^2\Big)^{(-1)}}\Big),$$

$$\mathbf{s}_{(d,j,n)} = \Big(\frac{\sum_{(d,j,n)\in\mathcal{R}_s}\langle\gamma_{s(i-1)}(k), \psi^s_{(d,j,n)}(k)\rangle}{\Big(\sum_{(d,j,n)\in\mathcal{R}_s}\big(\psi^s_{(d,j,n)}(k)\big)^2\Big)^{(-1)}}\Big),$$

where $\gamma_c(i-1)(k)$ and $\gamma_s(i-1)(h)$ $(k = 1, ..., N_{l_1}$ $h = 1, ..., N_{l_2})$ are the residuals of the stage $(i-1)$.

2. Let

$$\mathcal{V} = \sum_{k=1}^{N_{l_1}}\sum_{h=1}^{N_{l_2}}\Big(\frac{\hat{\mathcal{P}}\big(\gamma_{c(i-1)}(k)\big)}{\mathcal{P}\big(\gamma_{c(i-1)}(k)\big)}\Big)\ln\Big(\frac{\frac{\hat{\mathcal{P}}\big(\gamma_{c(i-1)}(k)\big)}{\mathcal{P}\big(\gamma_{c(i-1)}(k)\big)}}{\frac{\hat{\mathcal{P}}\big(\gamma_{s(i-1)}(h)\big)}{\mathcal{P}\big(\gamma_{s(i-1)}(h)\big)}}\Big),$$

where $\mathcal{P}\big(\gamma_{c(i-1)}(k)\big) = \|\gamma_{c(i-1)}(k)\|^2$ is the true probability and the

$$\hat{\mathcal{P}}\big(\gamma_{c(i-1)}(k)\big) = \|\sum_{(d,j,n)\in\mathcal{R}_c}\mathbf{c}_{(d,j,n)}(k)\psi^c_{(d,j,n)}(k)\|^2$$

is the estimated probability. The same way for the true probability $\mathcal{P}\big(\gamma_{s(i-1)}(h)\big) = \|\gamma_{s(i-1)}(h)\|^2$ and the

$$\hat{\mathcal{P}}\big(\gamma_{s(i-1)}(h)\big) = \|\sum_{(d,j,n)\in\mathcal{R}_s}\mathbf{s}_{(d,j,n)}(h)\psi^s_{(d,j,n)}(h)\|^2.$$

$$\arg\big(\max_{\{\mathcal{R}_c,\mathcal{R}_s\}}\|\mathcal{V}\|\big) = \{l_{c(d,j,n)}, l_{s(d,j,n)}\} \qquad (18)$$

with $(d,j,n) \in \{\mathcal{R}_c, \mathcal{R}_s\}$.

(This step selects the adaptive dilation on the cosine and sine frames).

3. Update $\mathbf{f}_c(t)$, γ_c, $\mathbf{f}_s(t)$ and γ_s:

$$\mathbf{f}_{c_i}(t) = \mathbf{f}_{c_{(i-1)}}(t) + \sum_{l_{c_{(d,j,n)}}\in\mathcal{R}_c}\mathbf{c}_{l_{c_{(d,j,n)}}}\psi^c_{l_{c_{(d,j,n)}}}(t)$$

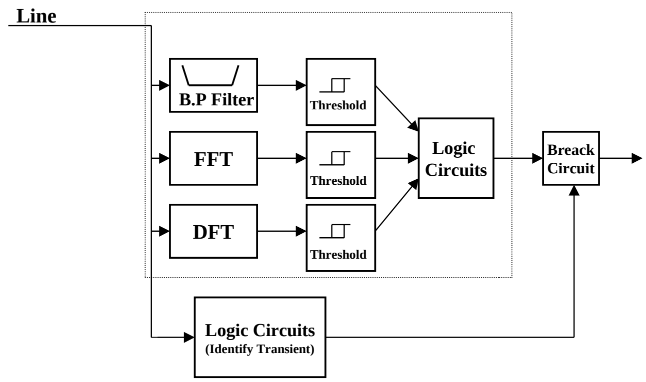

Figure 7: Actual Apparatus for Identification of Dangerous Transients.

$$\gamma_{c_i}(k) = \gamma_{c(i-1)}(k) - \sum_{l_{c_{(d,j,n)}}\in\mathcal{R}_c}\mathbf{c}_{l_{c_{(d,j,n)}}}(k)\psi^c_{l_{(d,j,n)}}(k);$$

$k = 1, ..., N_{l_1}$,

$$\mathbf{f}_{s_i}(t) = \mathbf{f}_{s_{(i-1)}}(t) + \sum_{l_{s_{(d,j,n)}}\in\mathcal{R}_s}\mathbf{s}_{l_{s_{(d,j,n)}}}\psi^s_{l_{s_{(d,j,n)}}}(t),$$

$$\gamma_{s_i}(h) = \gamma_{s(i-1)}(h) - \sum_{l_{s_{(d,j,n)}}\in\mathcal{R}_s}\mathbf{s}_{l_{s_{(d,j,n)}}}(h)\psi^s_{ls_{(d,j,n)}}(h);$$

$h = 1, ..., N_{l_2}$.

End Loop.

5 Simulations

In the preliminary simulation we have considered 20 signals as training signals and other 20 signals as fresh testing signals. These signals were split in two complementary classes, 12 signals with exceed dangerous limits around the bands 90 Hz and 105 Hz and 8 signals which don't exceed them.

It was known that the algorithm in [2] was able to recognize without any error the class which we have called $\mathcal{C}_3 = \{Inrush\ current\}$ performing adaptive subspaces on the *sine and cosine* wavelet packet trees organized in two levels.

The proposed algorithm generates, through the training, prototype signals and, after the compression through the same algorithm, coordinate systems so organized. A subspace $\mathcal{C}_3 = \{Inrush\ current\}$ characterized from a basis with 4 vectors, 2 (sine/cosine bases) for $\mathcal{C}_1 = \{No\ dangerous\ inrush\ current\}$ and 2 (sine/cosine bases) for $\mathcal{C}_1 = \{Dangerous\ inrush\ current\}$, every vector is characterized from 4 components corresponding to the [0; 50; 100; 150] Hz. The inner product between the coordinate systems and the fresh data vector resulting from the compression through the algorithm needs mostly only one time-frequency cell to recognize the inrush. The preliminary testing simulations, consisting of evaluating the Euclidean

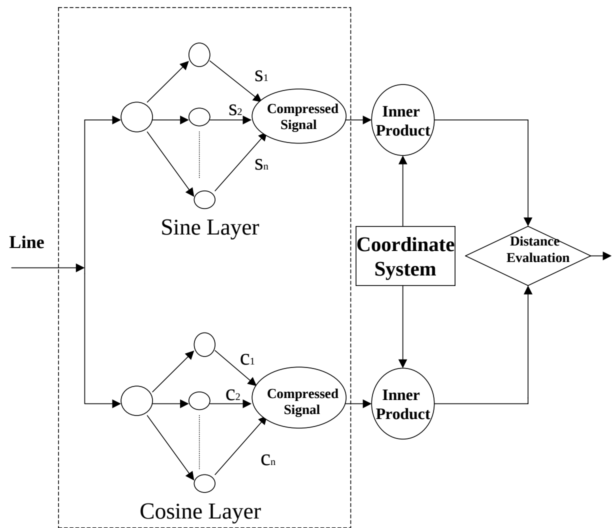

Figure 8: Neural Network Apparatus for Identification of Dangerous Transients.

distance between the projected fresh signal and the coordinate basis vectors, show correct identification percentages of 100 % in 20 ms. The classical methods used in rail vehicle control which combine FFT, DFT and bandpass plus threshold criterion needs normally more than 200 ms to recognize the inrush and their equipment are very tricky, see Fig. 7 and further information in [13]. In Fig. 8 is schematically reported the new network scheme for identification of dangerous transients, where it is possible to see its elementary structure.

6 Conclusions

We have described an algorithm to construct an adaptive orthonormal coordinate basis for classification problems of embedded classes of signals using wavelet packets in neural networks. The basis functions generated by this algorithm capture relevant time and frequency features in data. We proposed an algorithm for neural network training and filtering based on recursive iterations over biorthogonal library frames in order to detach, compress and classify the signals. The developed algorithm combines threshold techniques, regression analysis and backpropagation procedures to build a basis which *illuminates* the differences among embedded classes. It describes every complementary class with biorthogonal bases in order to perform an orthonormal coordinate system maximizing, at every step, the *Cross Entropy function*. In this way complementary phenomena are considered such as bimodal phenomena and the biorthogonal approximating bases allow to perform for each complementary class two orthonormal coordinate basis systems.

Acknowledgements

Paolo Mercorelli acknowledges financial support from E. C. through the Marie Curie Research Training Program.

References

[1] P. Mercorelli, M. Rode, and P. Terwiesch. *A wavelet packet algorithm for online detection of pantograph vibrations.* In Proc. 9 th IFAC Int. Symp. on Control in System Transportation 2000, in press, Braunschweig, 2000.

[2] P. Terwiesch and P. Mercorelli. *A signal classification algorithm using smooth local trigonometric bases.* Submitted to IEEE Transactions on Neural Networks. (Available via ftp in www.dii.unisi.it/˜mercorel/), 1999.

[3] P. Terwiesch, S. Menth, and S. Schmidt. *Analysis of transients in electrical railway networks using wavelets.* IEEE Trans. on Industrial Electronics, 45(6):955–959, 1998.

[4] Q. Zhang. *Using wavelet network in nonparametric estimation.* IEEE Transaction on Neural Networks, 8(2):227–236, 1997.

[5] D. L. Donoho. http://playfair.stanford.edu/˜wavelab. 1996.

[6] B. Jawerth and W. Sweldens. *Biorthogonal smooth local trigonometric bases.* J. Fourier Anal. Appl., 2(2):109–133, 1995.

[7] I. Daubechies. *Ten Lectures on Wavelets.* Publisher Society for Industrial and Applied Mathematics, Philadelphia (Pennsylvania), 1995.

[8] N. Saito and Ronald R. Coifman. *Local discriminant bases.* Mathematical Imaging: Wavelet Applications in Signal and Image Processing II, Proc. SPIE, 2303, 1994.

[9] A. Benveniste, A. Juditscky, B. Delyon, Q. Zhang, and P. G. Glorenee. *Wavelets in identification.* In Proc. SYSID'94 10th IFAC Symp. Syst. Identification, Copenhagen, 1994.

[10] P. Auscher, G. Weiss, and M. V. Wickerhauser. *Local sine and cosine bases of Coifman and Meyer and the construction of smooth wavelets.* Wavelet-A Tutorial in Theory and Applications, C-K Chui (ed.) ISBN 0-12-174590-2, Academic Press, Boston, pages 237–256, 1992.

[11] R. R Coifman and M. V. Wickerhauser. *Entropy based algorithm for best basis selection.* IEEE Trans. Inform. Theory, 32:712–718, 1992.

[12] Y. Meyer. *De la recherche pétrolière à la géometrie des espaces de Banach en passant par les paraproduits.* Preprint: École Polytechnique, Palaiseau, Seminaire équations aux dérivées partielles, 1986.

[13] www.enotrac.com/escarv/index.htm.