A Novel Criterion of Wavelet Packet Best Basis Selection for Signal Classification With Application to Brain–Computer Interfaces

Denis Vautrin, Xavier Artusi, Marie-Françoise Lucas, and Dario Farina*

***Abstract*—This study proposes a method to select a wavelet basis for classification. It uses a strategy defined by Wickerhauser and Coifman and proposes a new additive criterion describing the contrast between classes. Its performance is compared with other approaches on simulated signals and on experimental EEG signals for brain–computer interface applications.**

***Index Terms*—Brain-computer interfaces (BCIs), pattern recognition, wavelets.**

I. Introduction

A brain–computer interface (BCI) is a system that allows communication between the brain and a computer or a robot, without the use of nerves or muscles. The inputs of the system are usually EEG signals recorded on the scalp's surface. The output is a decision of action among a set of possible ones (for example, a command to a prosthesis). A training session is usually required to build the decision rules that allow the decoding of the user's intention [1].

The core of a BCI is a classification algorithm in which the EEG signals are mapped into the space of signal descriptors (feature space) and classified using decision functions learned on the training set composed of labeled signals. The performance of the classification depends on the choice of the feature space and on the type of discriminant functions. In this paper, our objective is to optimize the mapping (i.e. the transform applied to each signal) in order to obtain the best classification results. The energy distribution over uniform frequency subbands given by the Fourier transform is an example of *a priori* choice of signal features. In previous studies [4], [7], we have proposed the marginals of the discrete wavelet transform (DWT) for feature extraction and the feature space was selected by optimizing the mother wavelet of the decomposition. The DWT marginals reflect the average signal intensity over dyadic subbands. The dyadic decomposition is well suited to describe and discriminate signals whose discriminative information is mainly at low frequencies since the frequency resolution is higher for low frequencies than for high frequencies. However, in many cases, the frequency subbands that allow best discrimination are not known *a priori*.

Generalization of the DWT in the discrete wavelet packet transform (DWPT) allows a subband analysis without the constraint of a dyadic decomposition. The DWPT performs a redundant dichotomic decompositon of the frequency axis, but a specific decomposition corresponding to a particular basis may be selected according to an optimization criterion. In this study, we present a method for signal classification based on DWPT decomposition and optimization of the basis of wavelet packets. The method is based on a novel additive contrast criterion. The proposed technique allows the adaptation of the frequency subbands in order to best separate the different classes, and in this study, it is combined with a robust classifier (support vector machine, SVM). A learning set composed of labeled signals is used for first selecting the most discriminant wavelet packet basis (that determines the feature space) and then for learning the decision rules of the SVM classifier in this space (supervised classification).

We first present (see Section II) the feature space based on DWT and DWPT, the method to optimize it and the decision rule in the optimal feature space. The proposed technique can be applied to any classification problem. It was evaluated by comparison with two other approaches on synthetic signals and on experimental EEG signals in a representative BCI application (see Section III). The results are discussed in Section IV.

II. Methods

A. Feature Space

1) DWT and DWPT: In the context of the multiresolution analysis, the DWT is defined from a scaling function $\phi(t)$ and a mother wavelet $\psi(t)$, or equivalently by the corresponding scaling and wavelet (conjugate mirror) filters h and g. It performs a signal analysis according to a dyadic decomposition of the frequency axis. A DWPT [8] is a generalization of the DWT that performs a dichotomic and redundant exploration of the frequency content of a signal, which allows adaptation to particular signals. DWPT is briefly recalled here in the context of orthogonal wavelets. From a scaling function $\phi(t)$, the wavelets at the resolution level $j+1$ are defined as follows:

$$\psi_0^0(t) = \phi(t)$$

$$\begin{cases} \psi_{j+1}^{2p}(t) = \sum_k h[k]\psi_j^p(t-2^j k) \\ \psi_{j+1}^{2p+1}(t) = \sum_k g[k]\psi_j^p(t-2^j k) \end{cases}, \quad p = 0,\ldots,2^j-1. \quad (1)$$

Manuscript received March 6, 2009; revised May 26, 2009. First published July 31, 2009; current version published October 16, 2009. *Asterisk indicates corresponding author*.

D. Vautrin, X. Artusi, and M.-F. Lucas are with the Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN), Centrale Nantes, Nantes 44321, France.

*D. Farina is with the Center for Sensory–Motor Interaction, Department of Health Science and Technology, Aalborg University, Aalborg DK-9220, Denmark (e-mail: df@hst.aau.dk).

Digital Object Identifier 10.1109/TBME.2009.2028014

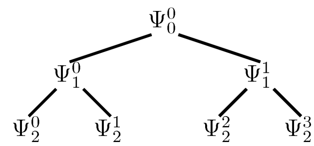

Fig. 1. Wavelet packets tree, with $J = 2$. The subbasis Ψ_j^p explores the normalized frequency band $2^{-(j+1)}[p, p+1]$.

These functions are organized according to a binary tree whose nodes (wavelet packets) are the subbasis $\Psi_j^p = \{\psi_j^p(t - 2^j k)\}_k$ with j defining the resolution level and p the analyzed frequency band at this resolution (see Fig. 1). The subbasis Ψ_j^p explores the normalized frequency band $2^{-(j+1)}[p, p+1]$. The whole library of wavelet packets is defined by $\{\{\Psi_j^p\}_{p=0,\ldots,2^j-1}\}_{j=0,\ldots,J}$.

The coefficients of the decomposition of a discrete-time signal x, of length N, are given by the generalization of the Mallat algorithm [8]

$$cx_0^0[k] = x[k], \qquad k = 0, \ldots, N-1$$

$$\begin{cases} cx_{j+1}^{2p}[k] = \downarrow 2[cx_j^p * \overline{h}][k] \\ cx_{j+1}^{2p+1}[k] = \downarrow 2[cx_j^p * \overline{g}][k] \end{cases}, \quad \begin{array}{l} p = 0, \ldots, 2^j - 1 \\ k = 0, \ldots, \dfrac{N}{2^{j+1}} - 1 \\ j+1 = 1, \ldots, J \end{array} \quad (2)$$

where $\downarrow 2$ corresponds to the downsampling operator $\overline{f}[k] = f[-k]$, and $J \leqslant \log_2(N)$ is the deepest level of the decomposition. Each packet of wavelet coefficients $\Psi_j^p x = \{cx_j^p[k]\}_{k=0,N/2^j-1}$ (where $cx_j^p[k] = \langle x, \psi_j^p(t - 2^j k)\rangle$) contains information on the signal in the particular frequency band explored by Ψ_j^p.

A basis can be extracted from the whole redundant set of wavelet packets by means of several strategies. A basis corresponds to a set of packets $\{\Psi_j^p\}$ that covers the frequency axis without overlap. This paper presents a new strategy for selecting such a basis with the objective of classification (see Section II-B). The decomposition of the signal x on a basis B is defined by

$$Bx = \{\Psi_j^p x | \Psi_j^p \in B\}. \quad (3)$$

The DWT performs the decomposition on a particular basis that is obtained by the restriction $p = 0$ in (1) and corresponds to $\{\{\Psi_j^1\}_{j=1,\ldots,J}, \Psi_J^0\}$.

2) Marginals: The wavelet packet coefficients can be processed before being used by the classifier. For example, in many applications, the analyzed signals are not perfectly time-aligned. As a consequence, the features used for the classification should only be composed of frequency descriptors. In this case, we have previously proposed to represent a signal by its marginals (that can be seen as signal intensity) over the frequency bands [4], [7]. They are calculated at each node of the decomposition and defined as follows:

$$M\Psi_j^p x = \sum_{k=0}^{N/2^j-1} |cx_j^p[k]|. \quad (4)$$

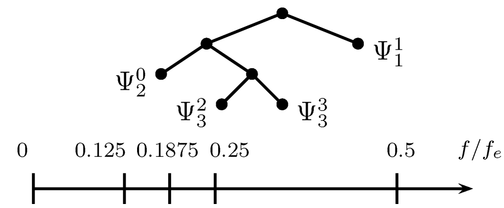

Fig. 2. Basis corresponds to the leaves of a pruned binary tree and explores the frequency axis without overlap. We show here an example of basis extracted from a DWPT tree ($J \geq 3$) and the corresponding normalized frequency bands (f_e: sampling frequency).

For a given basis B, the feature vector is given by

$$MBx = [\ldots\ M\Psi_j^p x\ \ldots], \qquad \text{where } \Psi_j^p \in B. \quad (5)$$

The use of marginals is an example of transformation of the wavelet coefficients; any other transformation can be applied with the best basis selection algorithm proposed in the following.

B. Optimization of the Feature Space

As discussed earlier, the whole set of wavelet packets is highly redundant: Ψ_j^p explores the same frequency interval as the union of its two sons $\Psi_{j+1}^{2p} \cup \Psi_{j+1}^{2p+1}$. In order to get a complete and nonredundant representation, we must select wavelet packets belonging to a basis, i.e., a set of packets $\{\Psi_j^p\}$ that covers the frequency axis without overlap. This corresponds to the leaves of an admissible tree, i.e., a tree in which each node has 0 or 2 sons. An example is given in Fig. 2 with the corresponding frequency bands.

To determine a basis adapted to the objective of classification, it is necessary to identify an appropriate selection criterion and a search strategy to optimize this criterion. We will use the search strategy proposed by Coifman and Wickerhauser [2], and propose a novel additive criterion, which will be described in the general context of multiclass decision.

1) Search Strategy of Wickerhauser and Coifman: Wickerhauser and Coifman have defined a local search strategy for getting a globally optimal basis from a library of basis having a binary tree structure [2]. Initially developed for adapting the basis for compression purposes, this algorithm can be associated to any objective under the condition that the criterion used is additive. A criterion $\mathcal{C}$ is said to be additive if, given A_1 and A_2 two disjoint subbasis, $\mathcal{C}(A_1 \cup A_2) = \mathcal{C}(A_1) + \mathcal{C}(A_2)$. The principle of the Coifman and Wickerhauser (CW) best basis search strategy is the following: starting from the bottom of the tree and going up till the root, one compares each father node to the union of its sons and keeps the better according to the selected criterion. The additivity of the criterion guarantees that the algorithm reaches the optimal basis and reduces the computational time.

2) Criterion: In the case of classification, the additive criterion should measure the discriminant capacity of any subbasis according to a learning population. Therefore, we search the basis that maximizes this criterion. We denote the learning set $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_i, \ldots, \mathcal{X}_c\}$, where $\mathcal{X}_i$ is composed of n_i

signals x belonging to the class ω_i ($i = 1, \ldots, c$), and n denotes the total number of signals ($n = \sum_{i=1}^{c} n_i$).

One additive criterion was proposed by Saito and Coifman [9] and consists in the distance (Euclidian or Kullback–Leibler distance) between the representatives of the classes. This criterion has been adapted for electromyographic (EMG) classification tasks [6]. It was originally designed to compare time-frequency representations. In this study, we first adapted the method by Saito and Coifman [9] to transformed coefficient (e.g., marginals) as follows. For a subbasis A, where each signal x of $\mathcal{X}$ is represented by its marginals vector MAx [as defined in (5)], the criterion is defined by

$$\mathcal{C}_1(A) = \sum_{i=1}^{c} \frac{n_i}{n} \|\overline{MA_i} - \overline{MA}\|^2 \quad (6)$$

where:

1) $\overline{MA_i} = \frac{1}{n_i} \sum_{x \in \mathcal{X}_i} MAx$ is the representative of class ω_i;
2) $\overline{MA} = \frac{1}{c} \sum_{i=1}^{c} \overline{MA_i}$ is the global center.

This criterion is additive but does not consider the class dispersion (intraclass inertia). In order to include the class dispersion information, we propose to use a Fisher-type criterion. In its classical form, a Fisher-type criterion is given by the ratio of interclass inertia over intraclass inertia, which corresponds, for a subbasis A, to

$$\mathcal{C}_F(A) = \frac{\sum_{i=1}^{c} \frac{n_i}{n} \|\overline{MA_i} - \overline{MA}\|^2}{\sum_{i=1}^{c} \frac{n_i}{n} \overline{\overline{MA_i}}} \quad (7)$$

where $\overline{\overline{MA_i}} = \frac{1}{n_i} \sum_{x \in \mathcal{X}_i} \|MAx - \overline{MA_i}\|^2$ is the inertia of ω_i.

Maximizing this criterion corresponds to maximizing the distance between representatives while minimizing the dispersion of each class. However, this criterion is not additive, thus it may lead to a nonoptimal basis if used with the CW best basis search strategy.

We propose a new additive criterion defined as

$$\mathcal{C}_2(A) = K \cdot \sum_{i=1}^{c} \frac{n_i}{n} \|\overline{MA_i} - \overline{MA}\|^2 - (1-K) \cdot \sum_{i=1}^{c} \frac{n_i}{n} \overline{\overline{MA_i}}. \quad (8)$$

If K is chosen between 0 and 1, the same properties of the classic Fisher criterion are maintained: by maximizing $\mathcal{C}_2$, we maximize the distance between representatives and minimize the inertia of classes. Moreover, the criterion is additive and thus appropriate for the use of the strategy proposed by Wickerhauser and Coifman.

The pertinence of the criterion $\mathcal{C}_1$, previously proposed, and of the novel criterion $\mathcal{C}_2$ were assessed on simulated signals (see Section II-D) and experimental signals (see Section II-E). The value of K was fixed based on empirical optimization on simulated signals (giving optimal values for $K \in [0.3; 0.7]$). The value of 0.5 was then used for both simulated and experimental signals.

C. Classification Rule and Evaluation of the Method

1) Multiclass Decision Rule: For the multiclass problem (c classes $\omega_i, i = 1, \ldots, c$), we use a one-versus-rest (OVR) procedure: for each class ω_i, we design a two-class linear SVM [3] to separate ω_i from the rest of the population. After training on a learning set, each two-class linear SVM provides a linear discriminant function $f_i, i = 1, \ldots, c$. Then, for a signal x described by MBx, in the optimized feature space corresponding to the best basis B, the final decision is taken from the c discriminant functions f_i

$$\text{“}x \text{ belongs to } \omega_j\text{” with } j = \underset{i=1,\ldots,c}{\text{Argmax}}\, f_i(MBx). \quad (9)$$

2) Performance Evaluation: The performance of the method has been evaluated from the misclassification rate calculated on a test set. For the experimental case, as the number of available signals is small, we used a leave-one-out procedure (crossvalidation) [5]. The simulated and experimental results will refer to two-class problems.

D. Simulated Signals

The synthetic signals were built such that each class had energy on a particular frequency band. To do so, we used a linear combination of atoms well localized in frequency: a set of atoms was associated to each class and two signals of the same class differed by random time location of atoms, which was defined by a Bernoulli distribution. The simulation of a signal x of length N is defined as follows:

$$x[k] = \sum_{m=1}^{Nv^i} \sum_{k_0=0}^{N/2^{j_m}-1} q_m^i[k_0] v_m^i[k - 2^{j_m} k_0]$$
$$k = 0, \ldots, N-1 \quad (10)$$

where the variables are defined as follows.

1) i denotes the class of signal x.
2) $\{v_m^i\}_{m=1,\ldots,Nv^i}$ is the set of atoms used to build the signals of class ω_i.
3) j_m corresponds to the frequency resolution of atom v_m^i.
4) q_m^i is a vector of length $N/2^{j_i}$ which terms take values 0 or 1 following a Bernoulli distribution.

Independent, white Gaussian noise was added to the signals.

One simulation consisted of two classes of signals, both separated into a learning set composed of 10 or 100 signals per class, and a test set composed of 1000 signals per class. For both classes, signals were of length $N = 128$ and the parameter of the Bernoulli distribution was set to 0.05. The SNR was varied between 2 and 10 dB.

E. Experimental Signals

Experimental EEG signals were recorded from five healthy men (age 22–30 years), who provided written informed consent before participation. The experimental procedures were approved by the local ethical committee of the Region North Jutland (Denmark) (N-20070001). The subjects performed imaginary right-sided isometric plantar flexion at two rates of torque development. The ballistic task was imagined as the reaching of 60% of the maximal voluntary torque as fast as possible, whereas the moderate task was defined as a 4-s linear torque increase from 0% to 60% of the maximal torque. Each task comprised 75 imaginary trials and the task execution was randomized between the two tasks. The subject was prompted for the task with auditory warnings (one or two beeps for the two tasks). These warnings were given every 15 s.

EEG activity was recorded using a 40-channel digital dc EEG amplifier (NuAmps, Neuroscan Laboratories) from the location Cz (10–20 system standard positions). Additionally, vertical and horizontal electrooculogram (EOG) signals were recorded with standard tin electrodes. The EEG signal was referenced to linked ear lobes ($A1$ and $A2$). EEG and EOG signals were amplified by a gain of 1338, low-pass filtered at 100 Hz, and sampled at 500 Hz using a 22-bit A/D converter. EMG activity of the right soleus and tibialis anterior muscles was recorded from surface Ag/AgCl electrodes (Neuroline 720, Ambu) using a custom-made amplification system (Aalborg University). EMG signals were amplified by a gain of 1000, band-pass filtered between 10 Hz and 1 kHz, and sampled at 2 kHz. EMG monitoring was used for indicating eventual unwanted muscle contractions during imaginary motor tasks. Trials were excluded from the analysis if the EMG activity exceeded by 25% the noise root mean square or if the peak-to-peak EOG amplitude exceeded 75 μV in one of the four EOG channels.

The misclassification rates obtained from experimental recordings were analyzed with Friedman's analysis of variance (ANOVA) with factor the method used (CW with C_1, CW with C_2, and DWT). Pairwise comparisons were performed with Wilcoxon-matched pairs test. Statistical significance was set to $P < 0.05$.

III. Results

Simulated and experimental signals were analyzed with the Daubechies-2 wavelet. It corresponds to an arbitrary choice among standard orthogonal wavelets.

A. Simulated Signals

Twenty simulations were performed in each condition of learning size and SNR, and results are reported in Table I as mean and standard deviation of the percentage of incorrectly classified signals, as calculated on the test set. The best results were obtained with the best basis given by the CW strategy using criterion $\mathcal{C}_2$ over the use of $\mathcal{C}_1$ or DWT basis.

B. Experimental EEG Signals

After discarding the trials with high EOG or EMG, as described earlier, the number of available signals per class varied between 25 and 48, depending on the subject. Table II shows the results for experimental recordings. As for simulated signals, the use of the criterion $\mathcal{C}_2$ led to better results than the use of criterion $\mathcal{C}_1$ or the DWT basis. In the experimental conditions, the criterion $\mathcal{C}_1$ performed worse than the DWT basis. The proposed criterion allowed the classification of the two imaginary tasks with an average misclassification error of 26%.

A Friedman ANOVA with factor the method used for feature extraction (CW with C_1, CW with C_2, and DWT) was significant ($P < 0.05$). The pairwise comparisons indicated that the classification with CW and C_1 (41% $\pm$ 7%) led to significantly greater misclassification rate than using CW and C_2 (26% $\pm$ 7%). On the contrary, the misclassification rates achieved by CW with C_1 (41% $\pm$ 7%) and DWT (34% $\pm$ 11%) (the two methods used for comparison with the proposed method) were not statistically different.

TABLE I
Misclassification Rate on Simulated Signals

Methods Learning size, SNR	**CW using $\mathcal{C}_1$** ($J = 3$)	**CW using $\mathcal{C}_2$** ($J = 3,\ K = 0.5$)	**DWT basis**
10, 10 dB	26 $\pm$ 15%	17 $\pm$ 3%	22 $\pm$ 3%
100, 10 dB	14 $\pm$ 2%	11 $\pm$ 1%	18 $\pm$ 1%
10, 2 dB	32 $\pm$ 13%	26 $\pm$ 5%	29 $\pm$ 5%
100, 2 dB	21 $\pm$ 7%	19 $\pm$ 2%	24 $\pm$ 1%

TABLE II
Misclassification Rate on Experimental EEG Signals

Methods	**CW using $\mathcal{C}_1$** ($J = 3$)	**CW using $\mathcal{C}_2$** ($J = 3,\ K = 0.5$)	**DWT basis**
Subject 1	48%	32%	34%
Subject 2	42%	18%	22%
Subject 3	29%	24%	23%
Subject 4	45%	23%	44%
Subject 5	40%	34%	46%
Mean	*41%*	*26%*	*34%*
St. Dev.	*7%*	*7%*	*11%*

IV. Discussion

In this study, we have proposed a pattern recognition approach for BCI applications based on DWPT and SVM. The feature extraction was optimized by a novel additive criterion of class separability.

The proposed criterion $\mathcal{C}_2$ in the CW strategy led to better results than the criterion $\mathcal{C}_1$, in both simulated and experimental signals. This reflects the importance of including both the interclass and the intraclass inertias for the choice of the best basis. Another criterion that takes into account the dispersion of signal features was proposed in [10]. It was based on the comparison of the probability density function of the classes. However, in several applications, as the one presented in this study, the number of training samples may not be sufficient to accurately estimate their probability density function without any *a priori* information. The criterion proposed in this study is based on only two measures extracted from the feature space in the training set.

From the simulations, it was observed that the misclassification rate increased with decreasing SNR and number of learning signals. These tendencies were expected and reflect the overfitting phenomenon: the classifier provides an efficient discriminant function if the learning signals are sufficiently numerous and not largely affected by noise. Moreover, the improvement due to $\mathcal{C}_2$ was more important when using small learning sets, which is particularly relevant in applications such as BCI.

The experimental signals provided a representative application for the proposed classification scheme to BCI. They were recorded during imaginary tasks performed at different speeds. It has been previously shown that these signals can be discriminated on a single-trial basis [7] and this result has been confirmed in the present study. The use of DWPT and the criterion $\mathcal{C}_2$ on the experimental signals was superior with respect to DWT in all subjects, except for subject 3 (see Table II). On average, the performance improved substantially with the proposed approach with respect to the other approaches compared, especially with respect to the previously proposed $\mathcal{C}_1$ criterion which led to poor results.

In conclusion, we have proposed a novel additive discriminant criterion for best basis selection in the DWPT framework. This new criterion showed improved classification performance on both simulated and experimental signals. The proposed pattern recognition method allowed discrimination of imaginary tasks corresponding to two speeds that may be a useful control strategy for BCI applications [7]. A future further improvement of the method may be the inclusion of the optimization of the mother wavelet, in addition to that of the wavelet packet, as proposed in [4] and [7] instead of using a standard wavelet.

References

[1] N. Birbaumer, A. R. Murguialday, and L. Cohen, "Brain-computer interface in paralysis," *Curr. Opin. Neurol.*, vol. 21, no. 6, pp. 634–638, Dec. 2008.

[2] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithm for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[4] O. F. Do Nascimento and D. Farina, "Movement-related cortical potentials allow discrimination of rate of torque development in imaginary isometric plantar flexion," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 11, pp. 2675–2678, Nov. 2008.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[6] K. Englehart, B. Hudgins, and P. A. Parker, "A wavelet based continuous classification scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 3, pp. 302–311, Mar. 2001.

[7] D. Farina, O. F. Do Nascimento, M. F. Lucas, and C. Doncarli, "Optimization of wavelets for classification of movement-related cortical potentials generated by variation of force-related parameters," *J. Neurosci. Methods*, vol. 162, no. 1, pp. 357–363, May 2007.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[9] N. Saito and R. R. Coifman, "Local discriminant bases," *Proc. SPIE, Mathematical Imaging: Wavelet Applications in Signal and Image Processing II*, A. F. Laine and M. A. Unser, Eds. 1994, vol. 2303, pp. 2–14.

[10] N. Saito and R. R. Coifman, "Improved local discriminant bases using empirical probability density estimation," in *Proc. 1996 Comput. Section Amer. Statist. Assoc.*, 1997, pp. 312–321.