31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Denoising of Multiscale/Multiresolution Structural Feature Dictionaries for Rapid Training of a Brain Computer Interface

Nuri Fırat İnce, Vijay Aditya Tadipatri, Fikri Göksu, Ahmed H. Tewfik

Abstract—**Multichannel neural activities such as EEG or ECoG in a brain computer interface can be classified with subset selection algorithms running on large feature dictionaries describing subject specific features in spectral, temporal and spatial domain. While providing high accuracies in classification, the subset selection techniques are associated with long training times due to the large feature set constructed from multichannel neural recordings. In this paper we study a novel denoising technique for reducing the dimensionality of the feature space which decreases the computational complexity of the subset selection step radically without causing any degradation in the final classification accuracy. The denoising procedure was based on the comparison of the energy in a particular time segment and in a given scale/level to the energy of the raw data. By setting denoising threshold a priori the algorithm removes those nodes which fail to capture the energy in the raw data in a given scale. We provide experimental studies towards the classification of motor imagery related multichannel ECoG recordings for a brain computer interface. The denoising procedure was able to reach the same classification accuracy without denoising and a computational complexity around 5 times smaller. We also note that in some cases the denoised procedure performed better classification.**

I. Introduction

In the last few years there has been a growing interest on the use of adaptive time-frequency analysis methods for the classification of neural activity for a brain computer interface (BCI)[1-5]. Potential applications include both the analysis of invasively and noninvasively recorded multichannel activity such as electrocorticogram (ECoG) [4, 5] and electroencephalogram (EEG) [1. 2, 4 and 6] respectively. In particular, motor imagery related neural patterns have been captured in the construction of BCIs where the subjects are asked to imagine a particular motor activity such as hand or foot movement. Associated oscillatory patterns which occur as short lasting amplitude attenuation and increase are extracted from central areas for classification and used as a control command or a feedback to the user. In this scheme investigating the neural activity in a few seconds accompanying and following motor imagery can be successfully implemented by time-frequency and time-scale methods. One can say that these methods enable the researcher to see the “big picture” since these methods do not put any bias towards a particular time point or a predefined frequency band. Sample time-frequency maps obtained with a short-time Fourier transform from C3/C4 electrode locations of the standard 10/20 system during right hand motor imagery are given in Fig. 1. Short lasting event related desynchronization (ERD) in alpha band and the following synchronization (ERS) in the beta band can be clearly observed in the time-frequency maps. This motivates the researchers on using the time-frequency methods for classification. Although the time-frequency and time-scale methods provide noticeable information about the time varying content of the neural activity, these procedures generally provide redundant information for both signal representation and classification purposes. In particular in machine learning, redundancy in feature space is associated with over-learning which deteriorate the generalization capacity of the classifier on the unseen test set [7]. Consequently feeding the whole time-frequency plane to a machine learning algorithm can not be recognized as an efficient and a feasible strategy due to over-learning and computational complexity problems. To this end regularization step is necessary to prevent the classifier from over-fitting the time frequency plane and reduce the computations required to extract features and train the system. Feature subset selection and subspace projection are two of the main methods used in this field for dimension reduction. While subspace methods like principal component analysis (PCA) can be applied to both redundant and complete representation of the processed signal, it is has a higher computational complexity. This is due to the feature space, which is the time frequency plane, has to be computed for each observation. In addition computing reliable estimates of principal components for the high dimensional time-frequency plane requires large sets of training samples which is difficult to observe in real life situations. Hence subset selection procedures seem more advantageous for dimensionality reduction. Since a small amount of feature space is retained for final classification BCIs based on subset selection procedures may have very low computational complexity. However selecting/finding informative features from the time-frequency plane is a quite demanding task and it is associated with long training phases extending to several hours, therefore limiting their applicability in real life

N. F. Ince is with the departments of Neuroscience & Electrical and Computer Engineering, Twin Cities, MN 55455 USA (e-mail: firat@umn.edu).

V. A. Tadipatri is with Electrical and Computer Engineering Department, Twin Cities, MN 55455 USA (e-mail: tadip001@umn.edu).

F. Goksu is with the Electrical and Computer Engineering Department, Twin Cities, MN 55455 USA (phone: 612-624-5285; e-mail: goks0002@ umn.edu).

A. H. Tewfik is with Electrical and Computer Engineering Department, Twin Cities, MN 55455 USA (e-mail: tewfik@umn.edu).

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

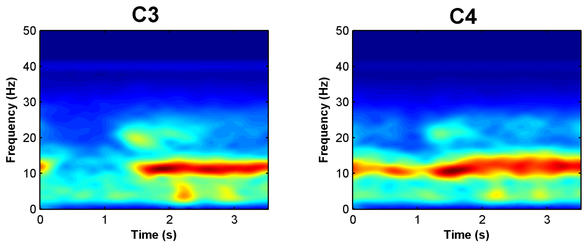

Fig1. The averaged EEG time frequency maps of C3 and C4 electrode locations during right hand finger movement imagery. On the C3 electrode location alpha band ERD is observable and it is followed by a beta band ERS around 20Hz.

scenarios.

In this study we propose a novel denoising step to reduce the computational complexity in the subset selection procedure causing dramatic shortening in the training time of a BCI. The paper is organized as follows. In the next section we shortly describe our previous studies towards structured feature generation and subset selection based on wavelets and block transform. In section III we describe the multichannel ECoG data set and provide experimental results. Finally, in section V we present our conclusion and future work.

II. Structured Feature Dictionary Generation and Subset Selection

Recently we developed new feature extraction and subset selection methods to address the issues highlighted above [4-6]. In particular we focused on two different main methods of constructing structural feature dictionaries based on (i) subband filtering and (ii) block transform of the neural activity. These methods rely on generating a redundant representation of the signal of interest and then capturing only necessary information from time-frequency plane by a subset selection procedure. A complete representation of the signal is not even necessary. In order to generate a redundant feature dictionary, in our previous study [4, 5], we used undecimated wavelet packet transform to decompose the neural activity into subbands in a tree structure and segmented each subband and used the energy in each time segment as a feature. Then we extended this method to block Fourier (BF) transform in [6]. Basically we interchanged the spectral and temporal segmentation processes. Following the feature generation a subset selection procedure evaluates a combination of features that originate from different spectral indices, temporal and/or spatial locations. The final selected subset of features is fed to the classifier for a decision. A block diagram representing the signal processing steps is given in Fig.2.a.

A. *Undecimated Wavelet Packet Transform*

The main structure in generating features is a dual tree as shown in Fig 2.b. Basically, we want to exploit the time and frequency content of the signal at a desired resolution. Therefore in our previous studies we decomposed the neural activity into subbands. Furthermore, each subband is divided into time segments. Both time and frequency segmentation steps use a dyadic tree structure. We used Undecimated Wavelet Packet Transform (UDWPT) to extract the subband activity. The undecimated wavelet transform is adapted to have the final transform to be shift invariant [8] which is crucial in pattern recognition applications. Once the segmentation is completed in time and frequency axis, the energy in each time segment, sum of coefficients' squares, is retained as one feature. This constitutes a feature space with dimension $N = (2^{(T+1)} – 1) (2^{(K+1)} – 1)$ where T is the time level and K is the frequency level. The reader is referred to [4] for details.

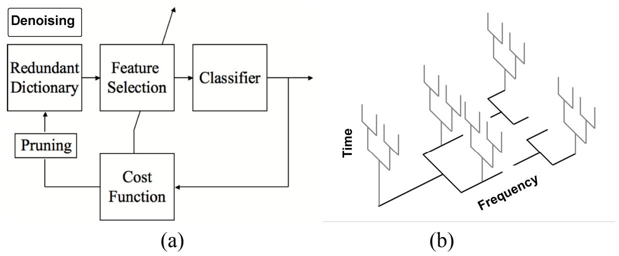

Fig2. (a) The block diagram of the proposed classification method. (b) The dual tree structure used to compute multiscale/multiresolution features.

B. *Block Fourier Transform*

We recently proposed to use the block fast Fourier transform (BF) as an alternative to the UDWPT to generate the features [6]. The features are generated using the same dual tree structure explained in the previous part. The only difference is that the subbands are extracted using harmonics. However, in this process the time segmentation precedes the frequency segmentation.

First, as explained in the previous section, the original signal is divided into $(2^{(T+1)} – 1)$ dyadic time blocks. Then, for each time block an N- point FFT is calculated, N being the length of the signal. Finally, each N-point FFT is segmented into $(2^{(K+1)} – 1)$ dyadic segments to extract the subbands. This is accomplished by merging consecutive expansion coefficients in desired subbands. The energy of each subband is retained as a feature. As in the previous case, there will be a total of $(2^{(T+1)} – 1) (2^{(K+1)} – 1)$ features.

C. *Sequential Subset Selection with Pruning (SFFS-P)*

The motivation in designing a feature dictionary using one of the methods explained in the previous section is to generate a redundant feature dictionary and let the classifier select a subset of them. This procedure implements a wrapper strategy to sequentially select features. In addition, SFFS-P executes a pruning method by using the structural relationship between dual tree nodes. In particular the pruning uses the information that all features have, which is a known location in time and frequency. When a feature is selected using the SFFS, the features that overlap in both the time and frequency are discarded. The SFFS algorithm is then applied on the remaining features. This reduces the computational complexity when compared to the stand alone sequential forward selection procedure.

The cost function to evaluate the discrimination ability of the combination of features is selected via the Fisher Discrimination (FD) criterion. In this study the FD measures how well two distributions are separated from each other in a one dimensional space after projecting them with linear discriminate analysis (LDA) [7].

D. Multiscale Denoising For Dimension Reduction

In a typical setting for 3-4 seconds of neural data we found that a time level of T=3 and a frequency level of K=4 is appropriate to extract redundant time-frequency or time-scale features. For a single channel this results in 465 features. For multichannel recordings this number reaches several thousands which in turn yields a very high computational complexity in the subset selection procedure. Recall that a pruning step was proposed by the authors in [4, 5] to reduce the computational complexity by removing overlapped time-frequency nodes. Since this procedure is executed on a channel basis, in case of a large number of channels, the pruning step has negligible effect on the speed of the training procedure. For instance in an EEG classification experiment we conducted in [6] the training time for 8 features over 33 channels and for T=3 and K=4 is around 3 hours.

Here we propose to use a denoising procedure to speed up the training step by removing a large portion of feature space a priori. Wavelets are successfully used for denoising audio and image signals. After computing the expansion coefficients they are squared and then sorted. Those values less than a threshold are removed. In the synthesis step only the ones exceeding the threshold are used for reconstruction. Although this procedure has proven its success in many compression scenarios we recognize it can not be applied directly to the motor imagery related neural activity. As indicated previously the rhythmic components of neural activity are modulated due to the motor imagery which appears as amplitude attenuation. Consequently those segments related to imagination appear as low energy segments and a straight forward application of a hard thresholding on the expansion coefficients may completely eliminate these events. In order to get around this problem we propose a novel denoising procedure which is based on the comparison of the energy in a particular time segment and at a given frequency scale to the energy of the raw data in this particular time segment. Here let $x_{t,k}^{c}$ be the average energy over trials, originating from channel $\boldsymbol{c}$ in time segment $\boldsymbol{t}$ and a frequency level $\boldsymbol{k}$, where $\boldsymbol{k}$=0 represent the raw neural data and $\boldsymbol{t}$=0 corresponds to the segment that covers the entire length of the signal**.** We define a variable, the relative energy ratio

$$E_{t,k}^{c} = \frac{x_{t,k}^{c}}{x_{t,0}^{c}} \tag{1}$$

in order to quantify how much of the variance of the raw data in a particular time segment ($x_{t,0}^{c}$) is captured in a given scale ($x_{t,k}^{c}$). It is not difficult to see that $E_{t,k}^{c} \in [0,1]$. Here the relative energy ratio enables us to indentify how much a localized feature contributes to the energy of that particular time segment and in a given scale/ frequency. Those values close to zero may be related to noise. Note that the normalization is implemented on a channel basis to eliminate the energy differences between different cortical areas. For instance, typically, in the occipital region large amplitude alpha waves are observed. In contrary, in the frontal areas, generally, slow waves such as alpha and theta bands are more dominant. Therefore we believe that the channel basis normalization suits more to the differences between different brain regions. After computing the $E_{t,k}$ for each node we sort them in a descending order and remove those feature indices smaller than a threshold, $\boldsymbol{d_{th}}$. The reader will immediately recognize that the sorting procedure can be executed simultaneously using the features from all channels. This provides a great opportunity in reducing the dimensionality of the feature space in a single step.

An important question here is how to define an appropriate denoising threshold. In this paper we studied the effect of denoising threshold $\boldsymbol{d_{th}}$ on the final classification accuracy and training time by setting it to three different values {0.01, 0.05, 0.1}. We executed several experiments on the following invasive multichannel neural recordings in order to asses the results of the denoising step.

E. Multi Channel EEG Data

We used the ECoG data set of BCI competition 2005 [9]. The ECoG data was recorded using an 8x8 ECoG platinum electrode grid which was placed on the contralateral (right) motor cortex. During the BCI experiment, a subject had to perform imagined movements of either the left small finger or the tongue. The channels of interest are converted to Hjort derivation in order to enhance the local activity [10]. Each channel was filtered with a low pass filter in 0-120Hz band. The filtered data was down sampled by a factor of 4 to 250Hz. Each trial was expanded from 750 samples to 768 samples by symmetric extension on the right side to enable segmentation in a pyramidal tree structure. We used 278 trials for training and 100 trials for testing. The training and test data were recorded from the same subject and with the same task, but on two different days with around 1 week in between. The challenge in this data set is to develop a robust classification system that can deal with intra-subject variability of the neural signals over time.

III. Results

In this study we set the time and frequency levels to 3 and 4 respectively when constructing the dual tree. Given the duration of the ECoG signal being 3 seconds with T=3, this corresponds to a finest time resolution of around 375ms (8 time segments). And for 125 Hz signal bandwidth with K=4, the finest frequency resolution was around 8 Hz. In the feature generation process using the dual tree we evaluated both UDWPT and block Fourier based methods. When the BF transform is used the basis functions are fixed. With the WP transform there are many filters to choose from. In order to see

TABLE I
THE CLASSIFICATION PERFORMANCE (%) OF THE DENOISING ALGORITHM AND THE NUMBER OF FEATURES (NOF) TO REACH MINIMAL ERROR WITH DIFFERENT WAVELET FILTERS IS PRESENTED.

db7	**UDWPT**		
	Test		**Percentage of Surviving**
d_{th}	**Error**	**NoF**	**Features**
0	7	3	100%
0.01	7	3	29.7%
0.05	**7**	**3**	**18.0%**
0.1	24	1	16.2%

db5	**UDWPT**		
	Test		**Percentage of Surviving**
d_{th}	**Error**	**NoF**	**Features**
0	7	3	100%
0.01	7	3	30.3%
0.05	8	5	18.1%
0.1	24	1	16.6%

the effect of vanishing moments on overall classification performance, two different wavelet filters are used; namely, 10 tap db5, 14 tap db7, respectively. Indeed higher number of vanishing moments provided a better frequency resolution with a cost of reduced temporal resolution.

We compared the classification accuracy and the training time of denoised cases to the case where no denoising was implemented. A 10-fold cross validation method was used to estimate the optimal model complexity (number of features) for each denoising level in the training set. Then these feature indices were used on the test set. These results for different wavelet filters and varying denoising levels are given in Tabel-I. We report the number of surviving features as a measure to quantify the complexity in the training stage since it is directly related to the number of features to be evaluated by the subset selection procedure. We note that the best classification accuracy with the least computational complexity was obtained with db7 wavelet filter at a denoising level of 0.05. In particular the training time was five times faster without causing any reduction in the final classification accuracy. Only three features were used for final classification step. We note similar characteristic on both db5 and db7 filters. At the denoising level 0.01 both setups reduced the training time with a scale of three. With both filters the algorithm selected the same features in alpha frequency band. The reader is referred to [5] for detailed information about the selected features.

TABLE II
THE CLASSIFICATION PERFORMANCE (%) OF THE DENOISING ALGORITHM AND THE NUMBER OF FEATURES (NOF) TO REACH MINIMAL ERROR IS PRESENTED.

	BLOCK - FFT		
	Test		**Percentage of Surviving**
d_{th}	**Error**	**NoF**	**Features**
0	11	2	100%
0.01	11	2	46.7%
0.05	10	4	26.0%
0.1	**10**	**4**	**20.6%**

We present the classification results obtained with block Fourier method in Table –II. In this particular setup we note that the denoising procedure not only decreased the computational complexity but also the error rate. For the denoising level of 0.1 the final classification error was 10% and the training time was improved almost 5 times. Only 4 features were used by the algorithm in the final classification step of the test data.

IV. CONCLUSIONS

In this study we extended the recently proposed dual tree structure based feature generation with a novel denoising procedure. The denoising procedure improved the training time radically without any loss in the final classification accuracy and in some cases with improved results. We note that the reduction in the number of candidate features was much larger in wavelet based approach which can be related to better energy packaging properties of this particular method [11]. We also note that the final classification accuracy of the wavelet based method is better than block transform based one. Here we believe that the better frequency resolution of wavelet filters with higher vanishing moments, db7, has an advantage over the block transform.

VI. REFERENCES

[1] T. Wang and B. He, “Classifying EEG-based motor imagery tasks by means of time–frequency synthesized spatial patterns”, Clin. Neuro. 115 2744, 2004.

[2] B. Yang, G. Yan, R. Yan and T. Wu, "Feature extraction for EEG-based brain–computer interfaces by wavelet packet best basis decomposition" J. Neural Eng. 3 251-256, 2006.

[3] N. F. Ince, S. Arica, and A. H. Tewfik, “Classification of single trial motor imagery EEG recordings by using subject adapted non-dyadic arbitrary time-frequency tilings,” J. Neural Eng. 3, 235-244, 2006.

[4] N. F. Ince, F. Goksu, and A. H. Tewfik, “An ECoG Based Brain Computer Interface with Spatially Adapted Time-Frequency Patterns,” Int. Joint. Conf. on Biomedical Engineering Systems and Technologies,, Portugal, 2008.

[5] N. F. Ince, F. Goksu, A. Tewfik, “ECoG Based Brain Computer Interface with Subset Selection”, Invited chapter, Communications in Computer and Information Science (CCIS) Book Series Springer, Biomedical Engineering Systems and Technologies,Vol25, 2008.

[6] F. Goksu, N. F. Ince, V. Tadipatri and A. H. Tewfik, “Classification of EEG with Structural Feature Dictionaries in a Brain Computer Interface”, in Proc. 2008 IEEE Engineering in Medicine and Biology Conf. (EMBC'08), Vancouver Canada, Aug. 2008.

[7] R. O. Duda, P. E. Hart, and D. G Stork, Pattern Classification. John Wiley & Sons, 2006, ch. 3.

[8] M. Unser, “Texture classification and segmentation using wavelet frames," IEEE Trans. Image Proc., Vol.4(11), pp.1549–60, 1995.

[9] http://ida.first.fraunhofer.de/projects/bci/competition_iii/

[10] B. Hjorth, “An online transformation of EEG scalp potentials into orthogonal source derivations,” Electroenceph. and Clinical Neuro. 39 526-30, 1975.

[11] S Mallat 2000 A wavelet tour of signal processing second edition Academic Press.