Pattern Recognition of Brain Signals

Pattern Recognition of Brain Signals

DISSERTATION

To obtain the degree of Doctor at the Maastricht University, on the authority of the Rector Magnificus, Prof. dr. G. P. M. F. Mols in accordance with the decision of the Board of Deans, to be defended in public on friday 24 of October 2008, at 12.00 hours

by

Federico De Martino

born on the 13 of March 1979 in Rome

Supervisor:

Prof. Dr. R. Goebel

Co-Supervisor:

dr. E. Formisano

Assessment Committee:

Prof. Dr. Eric O. Postma	*(Chairman)*
Prof. Dr. John-Dylan Haynes	*(Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany)*
Dr. Uri Hasson	*(New York University, New York, USA)*
Dr. Alard Roebroeck	*(University of Maastricht, Maastricht, The Netherlands)*

Contents

General Introduction

Based on:
Formisano, E., De Martino, F., Valente, G. (2007). Multivariate analysis of fMRI time series: classification and regression of brain responses using Machine Learning. Magnetic Resonance Imaging (*In Press*).

A wide range of multivariate statistical methods are being increasingly applied to the analysis of functional magnetic resonance imaging (fMRI) time series. These methodological developments are providing cognitive neuroscientists with the opportunity of tackling new research questions that are relevant for our understanding of the functional organization of the brain and that can often not be addressed using the widely employed univariate statistical methods. In this latter approach, a spatially invariant model of the expected blood oxygenation level dependent (BOLD) response is fitted independently at each voxel's time course and the differences between estimated activation levels during two or more experimental conditions are tested (Friston *et al.*, 1995). Together with methods for mitigating the problem of performing a large number of tests, this massively univariate analysis produces statistical maps of response differences, highlighting brain locations that are 'selective' or 'specialized' for a certain stimulus dimension, i.e. voxels or regions of interest (ROIs) that respond more vigorously to a sensory, motor or cognitive stimulus compared to one or more appropriate control conditions (Friston *et al.*, 1995). This analytical strategy focuses on the mapping of a 'stimulus – single location response' type of relation; however, it allows studying neither the stimulus (in-) dependent interactions between locations (functional and effective connectivity, see Friston *et al.*, 1994) nor the effects which are reflected in analytical strategies focusing on a 'stimulus – multiple locations response' type of relation.

Independent component analysis

Complementary to classical inferential analysis, 'massively multivariate' methods provide a characterization of the data, which does not rely on the statistical testing of a few stringent hypotheses and generate potentially valuable information on the nature of signal and noise in the fMRI time series. In some cases, however, the amount of information generated by these exploratory methods may be overwhelming and not easily interpretable.

When using spatial Independent Component Analysis (ICA), for example, fMRI time series are decomposed 'blindly' into a large number (up to the number of scans) of spatial modes (independent components, [ICs]), with associated time courses (McKeown *et al.*, 1998). These components reflect networks of 'functionally connected' (Friston *et al.*, 1994, McKeown *et al.*, 1998, Calhoun *et al.*, 2001, Beckmann *et al.*, 2004, Formisano *et al.*, 2004, Smolders *et al.*, 2007) brain areas. However, the order of extraction of ICs does not reflect stimulus specificity or neurophysiological relevance.

In chapter 1 of this thesis the issue of selecting 'interesting' and 'meaningful' independent components extracted from fMRI time series is addressed introducing the independent component fingerprint and applying pattern recognition methods to the classification of ICs.

In chapter 2 the application of the classification of ICs based on their fingerprint is shown in the context of focal epilepsy and the results of the selection of 'interesting' ICs are compared with conventional spike-related regression analysis.

Classification and regression of fMRI responses

Whereas a variety of methods for the investigation of functional (Friston *et al.*, 1994, McKeown *et al.*, 1998, Calhoun *et al.*, 2001, Beckmann *et al.*, 2004, Formisano *et al.*, 2004, Smolders *et al.*, 2007) and effective (e.g. McIntosh *et al.*, 1994, Friston *et al.*, 2003, Roebroeck *et al.*, 2005) connectivity have gained a conspicuous tradition in functional neuroimaging, only recently have methods been proposed for analysing the relation between a stimulus and the responses simultaneously measured at many locations (spatial response patterns or multivoxel response patterns).

This approach, which has been named multivoxel pattern analysis (MVPA, Norman *et al.*, 2006) or more evocatively 'brain reading' (see below), was initiated with a landmark fMRI study of the object-vision pathway (Haxby *et al.*, 2001). In this study, Haxby and colleagues demonstrated that spatial (multi-voxel) patterns of BOLD responses evoked by a visual stimulus are informative with respect to the perceptual or cognitive state of a subject. Participants were presented with various visual stimuli from different object categories (including faces, chairs, bottles); measured data were split in half and the spatial patterns of responses in the ventrotemporal cortex (the visual 'what' stream) were estimated for each category and for each half of the data. By comparing the spatial correlation between all patterns obtained from the first half of the data with those obtained from the second half of the data, Haxby and colleagues demonstrated that perceiving each object category was associated with a distinct spatial pattern of responses, and thus that these patterns could be used to 'decode' the perceptual or cognitive state of the subjects. These results did not change when the same analysis was performed after excluding the regions of maximal responses (e.g. after excluding the BOLD signals in the fusiform face area for the 'face' category). Importantly, these findings show that, information on the perceived 'object category' is entailed not only in the maximally responsive regions, but also in spatially wide and

distributed pattern of non maximal responses in the entire ventrotemporal cortex. Note that information in these latter responses is ignored in the conventional, subtraction based approach that is aimed at detecting voxel-by-voxel statistically significant activation level differences and thus only looks at the 'tip of the iceberg' of the overall information content carried by the measured response patterns.

Following the study by Haxby *et al.* (2001) several other groups examined, with increased methodological sophistication, the relation between sensory or cognitive stimuli and the spatial patterns of the measured response, and obtained remarkable results (Haynes *et al.*, 2005, Kamitani *et al.*, 2005, Cox *et al.*, 2003, Mitchell *et al.*, 2004, Mourao-Miranda *et al.*, 2005, Mourao-Miranda *et al.*, 2006, LaConte *et al.*, 2005, Kriegeskorte *et al.*, 2006). The methods employed derive mostly from statistical pattern recognition and machine learning and range from linear discriminant analysis (Haynes *et al.*, 2005, Kriegeskorte *et al.*, 2006), linear (Kamitani *et al.*, 2005, Mourao-Miranda *et al.*, 2005, Mourao-Miranda *et al.*, 2006, LaConte *et al.*, 2005) and non-linear (Cox *et al.*, 2003) support vector machine (SVM) and Gaussian naïve Bayes (GNB) classifiers (Mitchell *et al.*, 2004).

A major advantage of these methods compared to the conventional univariate statistical analysis is their increased sensitivity in discriminating perceptual and cognitive states. Statistical pattern recognition exploits and integrates the information available at many spatial locations, thus allowing the detection of perceptual and cognitive differences that may produce only weak single-voxel effects. This integration of information across multiple locations may range from considering only a small neighbourhood of adjacent spatial locations (locally multivariate analysis, see e.g. Kriegeskorte *et al.*, 2006) to jointly considering spatially remote regions or even voxels across the whole brain. Note that this integration of information is substantially different from spatial smoothing, commonly used in fMRI analysis. Spatial smoothing, indeed, increases sensitivity only when two conditions differ in terms of their regional mean activation levels. In these cases, the signal differences in the local neighbourhood of a position of interest are all (ideal case) in the same direction and are enhanced by spatial averaging. Conversely, if two conditions differ in terms of their fine-grained spatial activation patterns, spatial smoothing has a destructive effect and cancels out the discriminative information, which can be detected by pattern recognition methods.

The application of pattern recognition algorithms to the analysis of fMRI data has focused mainly on detecting brain areas that could reliably detect the presented stimulus as belonging to one of several classes (classification). Only recently (PBAIC 2006 competition), machine learning has been applied to the learning and prediction of a functional relationship between brain response patterns and a perceptual, cognitive or behavioural state of a subject that can be

expressed in terms of a continuous label (machine learning based regression). In what follows a joint formulation of the problems of classification and regression of fMRI responses is given and the general framework for the application of pattern recognition algorithms to the analysis of fMRI data is introduced.

Problem formulation

Consider **D**, **D'** as two data sets from a generic fMRI experiment, and **t**, **t'** as the labels that describe the sensory, motor or cognitive stimulation associated with **D** and **D'** respectively. Pattern analysis algorithms aim at "learning" a functional relationship between data **D** (*training dataset)* and label **t** in order to predict the unseen labels **t'** from the new dataset **D'** (*test dataset*). During this prediction, the algorithm blindly decodes the brain activation **D'** into the corresponding stimulus condition or cognitive state of the subject (as described by the prediction $\hat{\mathbf{t}}'$), hence the definition of 'brain reading'.

More formally, the problem that pattern analysis algorithms try to solve can be described as the learning of a function:

$$f = f(\mathbf{D},\mathbf{t},\theta) \qquad (1)$$

where **θ** denotes the set of adjustable model parameters that may be estimated during the training phase. The estimated function can be used on a new dataset to predict unseen labels:

$$\hat{\mathbf{t}}' = f(\mathbf{D}', \mathbf{D},\mathbf{t},\theta) \qquad (2)$$

where $\hat{\mathbf{t}}'$ denotes an estimate of the labels **t'**. Labels may assume discrete values and describe the stimulus or (assigned) subject's cognitive state during the acquisition of a set of images; for instance in an experiment with two conditions (blocked or event-related design), labels will be $t_i = +1$ for all images collected under condition 1 and $t_i = -1$ for all images collected under condition 2. As described in chapters 1-4, this prediction problem is referred to as 'classification'. In other cases, the stimulus or subject assigned state may be described using labels with continuous values. As described in chapter 5, the prediction problem is then referred to as 'regression'.

Figure 1 summarizes in a block diagram the steps that are generally performed in applying pattern recognition algorithms in fMRI. In the training phase the raw fMRI data are preprocessed in order to reduce the effects of noise. This step can be usually performed with well-known functional neuroimaging analysis software tools (e.g. AFNI, BrainVoyager, FSL, SPM). Prior to model training, the second step, is the selection of *relevant* features from the data. This step aims at reducing the complexity of the dataset and increasing the capabilities of the prediction scheme. During *model training*, several processing and feature extraction/selection cycles may be explored; the learning phase is therefore depicted

with feedback, indicating that pre-processing, feature extraction and model estimation cannot be regarded as completely separated parts of the training phase). Once the training phase is completed, a set of pre-processing parameters, a set of features and a set of model parameters is available, and the prediction is performed using the trained model on a new dataset after preprocessing with the same parameters as used for the training data .

Performance metrics, cross-validation and model selection

Together with the model (1), a suitable performance metric has to be defined. This performance indicates how good the classification/regression is, and it can be expressed as an error (or loss) function ε. In very general words, the aim of a prediction algorithm is to learn a model on the training dataset **D** that gives the minimum error on an unseen dataset test **D'**. The learning algorithm, thus, does not focus on finding the best model explaining the training data set **D**, but on mod-

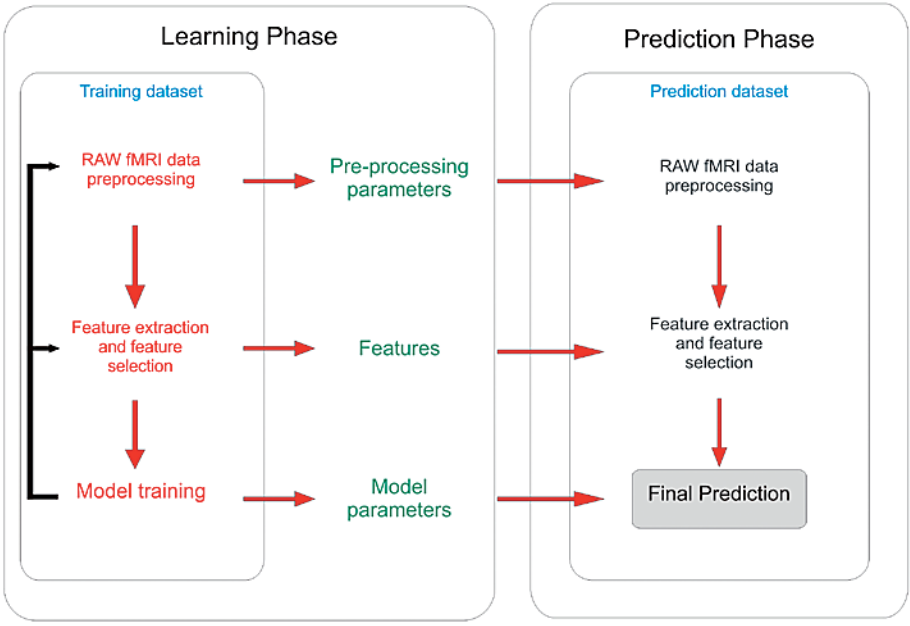

Figure 1. Main steps of a generic pattern recognition algorithm as used in fMRI data analysis. In the training phase (left) the raw fMRI data are preprocessed and relevant features are selected from the data prior to model training. Prediction (right) is performed using the trained model on a new dataset, after this latter has been preprocessed in the same way and reduced to same features as in the training.

els exhibiting the best generalization performance. Several models and several performance metrics can be introduced to perform this operation (Duda *et. al.*, 2001), some of which will be presented in the following chapters.

There are several ways to find suitable values for the model parameters, given the training dataset. One of the most employed approaches is to split the training dataset in two disjoint parts. One of the two is used as the traditional training set, while the other is used as a *validation* dataset, to evaluate the generalization error. Sometimes the dataset is divided into m disjoint parts and a model is trained m times, each time leaving a part as a validation set. The estimated performance is then defined as the mean of the errors of the models on the different datasets (*m-fold cross-validation*) (Duda *et. al.*, 2001).

Bayesian methods usually do not require cross-validation to compare different model parameter choices (although sometimes it is also used in Bayesian frameworks for model selection, see Rasmussen *et. al.*, 2006). In Bayesian analysis, model comparison involves the use of probabilities of the choice of a suitable model (Duda *et. al.*, 2001, Bishop 2006). Given a set of models $i = 1..., L$ that we wish to compare using the observed data $\mathbf{D}$ then it is possible to compare these models by means of their posterior distribution:

$$p(M_i|\mathbf{D}) = \frac{P(\mathbf{D}|M_i)P(M_i)}{P(\mathbf{D})} \propto P(\mathbf{D}|M_i)P(M_i) \qquad (3)$$

where the data-dependent term $P(\mathbf{D}|M_i)$, also known as model evidence, can be seen as a likelihood function over the models space and $P(M_i)$ denotes the prior probability of the model. The normalizing factor $P(\mathbf{D})$ can be discarded while comparing models. Once the posterior probability in (3) has been estimated, it can be used in two ways: 1) considering a linear combination of all the models, weighted by their probability (*mixture distribution*); 2) performing *model selection*, that is selecting the most probable model, given the data.

Chapter 3 addresses the relevant issue of voxel selection for the application of pattern recognition algorithms to fMRI data. Recursive Feature Elimination (RFE) is introduced for the selection of discriminative fMRI patterns and compared to previously applied feature selection strategies. This approach, starting from the whole brain, allows to iteratively refine and 'map' the most discriminative patterns. Chapter 4 describes the application of RFE in the context of an event-related auditory fMRI experiment aimed at deciphering activation patterns evoked by speech sounds into their content (vowels) or into the identity of the speaker.

In chapter 5, multivariate regression using pattern recognition algorithms is applied to the investigation of the coupling between electroencephalography (EEG) power modulations and fMRI BOLD signal. Regularized learning schemes are used to predict EEG power oscillations using patterns of simultaneously acquired fMRI data, and are compared to existing methods.

References

Beckmann, C.F., Smith, S.M. (2004). Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. IEEE Trans. on Medical Imaging;23(2):137-152.

Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J. (2001). Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Hum Brain Mapping;13(1):43-53.

Cox, D., Savoy, R. (2003). Functional magnetic resonance (fMRI) "Brain Reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage;19(2):261-270.

Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.

Duda, R.O., Hart, P.E., Stork, D.G. (2001). Pattern Classification. John Wiley & Sons, 2nd edition.

Friston, K.J. (1994). Functional and Effective connectivity in Neuroimaging: a Synthesis. Human Brain Mapping;2:56-78.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J. (1995). Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping;2:189-210.

Friston, K.J., Harrison, L., Penny, W. (2003). Dynamic Causal Modelling. NeuroImage;19(4):1273–1302.

Formisano, E., Esposito, F., Di Salle, F., Goebel, R. (2004). Cortex-based independent component analysis of fMRI time series. Human Brain Mapping;22(10): 1493-1504.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Aschouten, J.L., Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science;293(5539):2425-2430.

Haynes, J.D., Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci.;8(5):686-91.

Haynes, J.D., Rees, G. (2006). Decoding mental states from brain activity in humans. Nat. Rev. Neurosci.;7(7):523-34.

Kamitani, Y., Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci.;8(5):679-85.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America;103(10):3863-3868.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. Neuroimage;26(2):317-329.

McIntosh, A.R., Gonzalez-Lima, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. Human Brain Mapping;2:2–22.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T., Kindermann, S.S., Bell, A.J., Sejnowski, T.J. (1998). Analysis of fMRI data by blind source separation into independent spatial components. Human Brain Mapping;6:160-188.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X. (2004). Learning to decode cognitive states from brain images. Machine Learning;57:145-175.

Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage;28(4):980-95.

Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. Neuroimage;33(4):1055-65.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci.;10(9):424-30.

Rasmussen, C.E., Williams, C.K.I. (2006). Gaussian Processes for Machine Learning. MIT Press.

Roebroeck, A., Formisano, E., Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage;25:230-242.

Smolders, A., De Martino, F., Staëren, N., Scheunders, P., Sijbers, J., Goebel, R., Formisano, E. (2007). Dissecting cognitive stages with time-resolved fMRI data: a comparison of fuzzy clustering and independent component analysis. Magentic Resonance Imaging;25:860-868.

Classification of fMRI independent components *fingerprints*. 1

We present a general method for the classification of Independent Components (ICs) extracted from functional MRI (fMRI) data sets. The method consists of two steps. In the first step, each fMRI-IC is associated with an IC-*fingerprint*, i.e. a representation of the component in a multidimensonal space of parameters. These parameters are post-hoc estimates of global properties of the ICs and are largely independent of a specific experimental design and stimulus timing. In the second step a machine learning algorithm automatically separates the IC-*fingerprints* into six general classes after preliminary training performed on a small subset of expert-labeled components.

We illustrate this approach in a multi subject fMRI study employing visual structure-from-motion stimuli encoding faces and control random shapes. We show that: 1) IC-*fingerprints* are a valuable tool for the inspection, characterization and selection of fMRI-ICs and 2) automatic classifications of fMRI-ICs in new subjects present a high correspondence with those obtained by expert visual inspection of the components.

Importantly, our classification procedure highlights several neurophysiologically interesting processes. The most intriguing of which is reflected, with high intra- and inter-subject reproducibility, in one IC exhibiting a transiently task-related activation in the 'face' region of the primary sensorimotor cortex. This suggests that in addition to or as part of the mirror system, somatotopic regions of the sensorimotor cortex are involved in disambiguating the perception of a moving body part.

Finally, we show that the same classification algorithm can be successfully applied, without re-training, to fMRI collected using acquisition parameters, stimulation modality and timing considerably different from those used for training.

Based on:
De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., Formisano, E. (2006). Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. NeuroImage;34(1):177– 94.

Introduction

Non-inferential or exploratory multivariate methods are being increasingly used in fMRI data analysis. These methods provide a characterization of the data, which does not rely on the statistical testing of a few stringent hypotheses and generate potentially valuable information on the nature of signal and noise in the fMRI time series. The value of such information consists in being complementary to that of statistical inferential maps. In some cases, however, the amount of information generated by these exploratory methods may be overwhelming and not easily interpretable.

In spatial Independent Component Analysis (ICA), for example, fMRI time series are decomposed into a large number (up to the number of scans) of spatial modes (independent components, [ICs]), with associated time courses (McKeown *et. al.,* 1998). In most cases, some of these components reflect interesting spatiotemporal patterns of stimulus-induced or spontaneous brain activity; other components reflect signal artifacts or noise (McKeown *et. al.,* 2003). The basic assumption in spatial ICA is that fMRI time series can be modelled as linear mixtures of latent sources, which can be blindly recovered under the constraint that their spatial distributions are mutually statistically independent. Several recent methodological and applied contributions indicate that this approach outperforms Principal Component Analysis (PCA) and can be a useful complement to standard hypothesis-driven univariate analysis (McKeown *et. al.,* 1998). In contrast to PCA, however, in which extracted components are naturally ordered according to explained variance, ICA does not provide any intrinsic order of the ICs. The experimenter is thus confronted with the problem of selecting and interpreting a subset of 'interesting' and 'meaningful' components.

In previous fMRI applications of ICA, selection of interesting components has been performed using various approaches. The simplest approach relies on the visual inspection of IC- spatial maps/time courses (Bartels *et. al.*, 2005, Calhoun *et. al.*, 2001). Selection of ICs based on their visual inspection, however, is very time consuming and highly dependent on the experience of the researcher. In most cases, ICs have been selected according to the amount of linear correlation of their time course with a model of the expected responses (McKeown *et. al.,* 1998, Schmithorst *et. al.*, 2004, Moritz *et. al.*, 2005) or related measures in the temporal frequency domain (Moritz *et. al.,* 2003). These approaches, however, appear to contrast with the data-driven nature of ICA. As an explorative tool, ICA may be particularly useful for detecting patterns of activity whose temporal dynamics cannot be easily modelled, such as in the case of hallucinations (van de Ven *et. al.,* 2005), epileptic seizures or in sensory or cognitive paradigms in

which expected hemodynamic responses may be very diverse (Duann *et. al.*, 2002, Formisano *et. al.,* 2004, Castelo-Branco *et. al.,* 2002). Furthermore, ICA is being increasingly used for the study of 'resting state' functional connectivity (van de Ven *et. al.*, 2004, Greicius *et. al.*, 2003; 2004) or as a de-noising step, which requires the selection of components reflecting noise and artifacts (Thomas *et. al.*, 2002). In all these cases, selection of ICs based on strong expectations on the profile of the IC-time courses is insufficient.

Alternatively, selection of ICs has been performed using strong a priori assumptions on the spatial layout of the activation (Castelo-Branco *et. al.,* 2002; van de Ven *et. al.,* 2004). In this approach, distributed brain networks are detected by selecting ICA components that load heavily in pre-defined regions of interest (ROIs). A priori expectation on one or more ROIs, however, is not always available and, as in ROI-based univariate analysis, interesting processes occurring outside the pre-defined ROIs are ignored.

Other post-hoc measures obtained from estimated ICs have been used for their sorting/selection. In analogy to PCA, McKeown *et. al.* (1998) sorted the ICs according to their variance contribution to the original mixture. In fMRI data, however, neurophysiologically interesting phenomena are usually weaker than some of the sources representing structured noise. Thus, ranking of the ICs in this way may be not informative. Formisano *et. al.* (2002) characterized the ICs using a combination of three descriptive measures (kurtosis of the spatial distribution, one-lag autocorrelation of the IC-time course and clustering of the IC's spatial layout). The underlying idea was that 'meaningful' components aggregate in clustered regions in the three-dimensional space defined by these three measures. This heuristic criterion proved to be effective in isolating task-related components in a simple paradigm without using stimulus timing information.

In this chapter we introduce the IC-*fingerprint*, a visual tool that aids the experimenter in displaying and characterizing the ICs. An IC-*fingerprint* is a representation of the component in a multidimensional space of descriptive measures, which can be visualized as a polar diagram. In line with Formisano *et. al.* (2002), the underlying assumption is that ICs reflecting similar process types (e.g. BOLD activation, structured noise, movement) have similar *fingerprints.* To preserve the data-driven nature of ICA and the generality of the approach, the descriptive measures that define the space of the *fingerprints* are *post hoc* estimates of global properties of the ICs and do not rely on strong temporal or spatial hypotheses. The combination of ICA decomposition and IC-*fingeprint* characterization can be seen as "feature extraction steps" in the context of pattern recognition analysis (see figure 1 Introduction).

Furthermore we formulate the problem of selecting 'meaningful' components

in the more general context of their (automatic) classification. After transforming the ICs in the multidimensional space of *fingerprints*, this problem can be formulated as subdividing the ICs in maximally disjoint classes and finding the optimal separating set of boundaries (hypersurfaces). Many different (supervised and unsupervised) algorithms may be used for this purpose (see Mitchell, 1997). Here, we describe and validate a supervised method for the classification of the ICs based on least squares Support Vector Machines (ls-SVMs). SVMs refer to a class of machine learning algorithms introduced by Vapnik at the end of the 70s (Vapnik, 1979). Ls-SVMs are a variant of SVM which have been proved to be effective in many problems of classification and pattern recognition (Suykens *et. al.*, 2002).

We illustrate our approach in the context of a multisubject fMRI study with visual structure-from-motion (SFM) stimuli (Kriegeskorte *et. al.*, 2003). We show that the set of measures that defines the IC-*fingerprints* is informative with respect to the problem of selecting and classifying fMRI-ICs and allows a reliable detection of interesting activation patterns. Furthermore, we show that an ls-SVM classifier, which is trained with a small subset of data from one subject, can automatically classify these fMRI-ICs in all other subjects with high correspondence to an expert classification. Finally, we show that the same classification algorithm can be successfully applied, without re-training, to fMRI collected using magnetic field, acquisition parameters, stimulation modality (auditory vs visual) and timing (event-related vs block design) considerably different from the SFM experiment used for training.

Methods

General description of the approach

Figure 1 illustrates schematically the proposed approach. Individual fMRI time series are decomposed using spatial ICA in sets of ICs (see *ICA decomposition*). Obtained ICs are 'transformed' in a multidimensional space using a number of descriptive parameters (eleven in the current implementation). These parameters are computed from spatial distributions, spatial maps and time courses of ICs and characterize the global statistical and spatio-temporal nature of the sources (see *Characterization of the ICs in a multidimensional space*). We define IC-*fingerprint* as the representation of an IC in this multidimensional space of parameters (Figure 1a).

In order to classify the ICs, two steps are required (Figure 1b). The first

step consists in *training* the SVM classifier(s) using a limited subset of IC-*fingerprints* (typically the subset of ICs corresponding to an individual single run time series). This step requires expert user's intervention and supervision in labelling the ICs (see *Supervised training*). The second step operates fully automatically and consists in *classifying* all other components using the trained classifier(s) (see *Automatic classification of ICs*). Figure 1c visualizes, for the application presented in this article, the proportion of data which has been used for training (red)

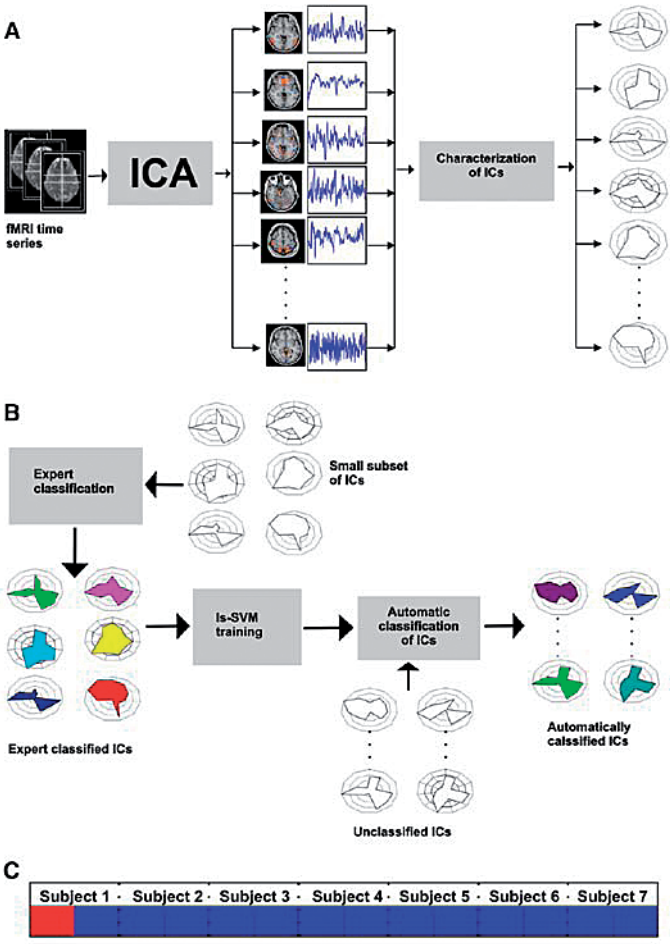

Figure 1: General description of the approach for the characterization and classification of fMRI-ICs; a) Independent Component Analysis of fMRI data and representation of the ICs in a multi dimensional space of fingerprints (see also Figure 2). b) Classification of IC-fingerprints by an ls-SVM based algorithm. The algorithm is trained on a small subset of data labeled by an expert. c) Proportion of data which has been used for training (red, 1/14) and testing (blue, 13/14) the classifier presented in this chapter.

and testing (blue) of the classification procedure.

ICA decomposition

Let **X** be the $T \times M$ (T = number of scans, M = number of time courses) matrix of the fMRI time series, **S** the $N \times M$ matrix whose rows S_i (i=1,...,N) contain the spatial processes ($N \leq T$ = number of processes) and **A** the $T \times N$ mixing matrix whose columns A_j (j=1,...,N) contain the time courses of the N processes and is assumed to be of full rank. The problem of the ICA-decomposition of fMRI time series can be formulated as the estimation of both matrices of the right side of the following equation:

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad (1)$$

under the constraint that the processes S_i are (in the ideal case) spatially independent. No a priori assumption is made about the mixing matrix **A**, i.e. about the time courses of the processes. In this model, all the spatial components, with the possible exception of one, are assumed to be non-Gaussian. Structured (non-Gaussian) artifacts in the data (e.g. head movements, machine and physiological artifacts) are not explicitly modelled, but instead are treated as independent sources and are expected to be represented in one or more of the components.

The amount of statistical dependence within a fixed number of spatial components can be quantified by means of their mutual information. Thus, the ICA decomposition of **X** can be defined as a linear transformation:

$$\mathbf{S} = \mathbf{W} \cdot \mathbf{X} \qquad (2)$$

where the matrix **W** (the "unmixing" matrix) is determined such that the mutual information of the target components S_i is minimized. Matrix **A** can be computed as the pseudo-inverse of **W**. Note that this definition of ICA and Eq. (2) implies that ICs are determined up to a permutation, a multiplicative constant and to the sign. In the present context, the indeterminacy with respect to the permutation is important because it implies that there is no intrinsic ordering of the ICs, which is in contrast with PCA.

We estimated **S** using cortex based ICA (cb-ICA) (Formisano *et. al.*, 2004) as implemented in BrainVoyager 2000 (Brain Innovation, Maastricht, The Netherlands). Cb-ICA uses individual anatomical constraints and a fixed point ICA algorithm (FastICA, see Hyvärinen, 1999) and allows an optimized analysis of cortical sources. After sphering the matrix **X** and reduction of the temporal dimension of the data set with PCA (see below), the hierarchical (deflation) mode of the FastICA algorithm was used and the components were estimated one-by-one. After the decomposition, voxel values of IC spatial maps were z-transformed and colour coded according to the absolute value and sign (McKeown *et. al.*, 1998). It should be noted that the z-values do not pertain to any significance statistic,

because no comparison is made to a null hypothesis.

Characterization of the ICs in a multidimensional space

Spatial ICA decompositions of fMRI time series result in sets of ICs, whose dimensionality is determined by the number of scans or by the (optional) reduction of the temporal dimensions with PCA. Conventionally, interpretation of results requires expert inspection of each IC's voxel values distribution (histogram) (Figure 2a), spatial map (Figure 2b) and time course (Figure 2c). Temporal information can be additionally expressed in the frequency domain by computation of a power spectrum (Figure 2d). The characterization of ICs which is described in this section has the purpose of aiding in this task and to highlight similarities among the ICs.

For each IC, values of eleven descriptive measures were derived from the IC's voxel values distribution (kurtosis, skewness, entropy), spatial layout (degree of clustering in the anatomical space) as well as their temporal (one-lag au-

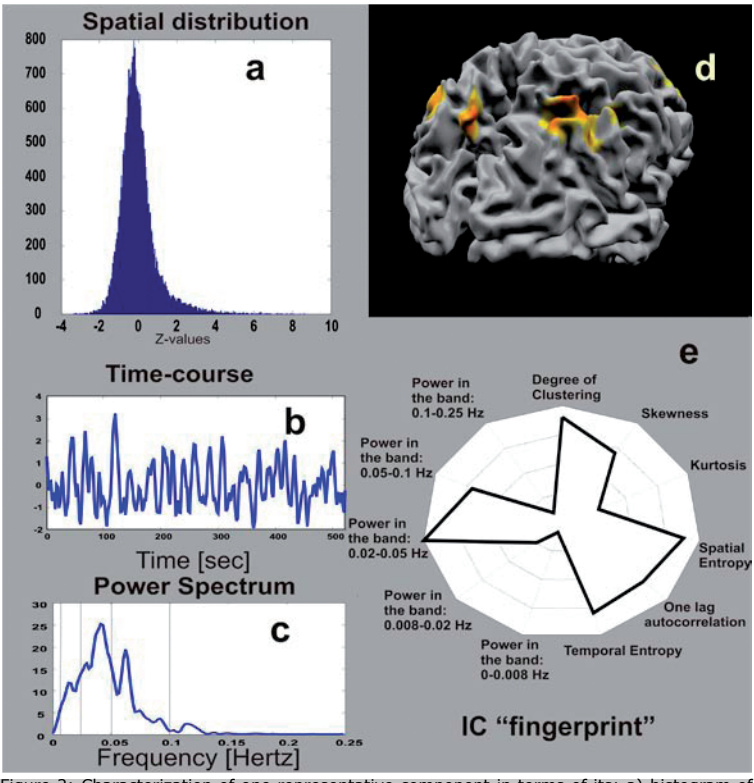

Figure 2: Characterization of one representative component in terms of its: a) histogram of voxel values; b) map layout (projected on the reconstructed cortical surface of the subject); c) time course; d) power spectrum; e) IC-*fingerprint*. Each axis in the polar plot corresponds to one of the normalized spatial, temporal or spectral parameters (see Appendix 1 for details).

tocorrelation, entropy) and spectral (power contribution in five different frequency-bands) properties. The exact definition of these measures and the rationale behind their inclusion are given in Appendix 1.

We define IC-*fingerprint* the representation of an IC as a point of the multidimensional (eleven-dimensional) space of parameters. IC-*fingerprints* are visualized using polar plots with eleven axes, each of them corresponding to one of the parameters normalized between 0 and 1 (Figure 2e).

Classification of the ICs with least-squares support vector machines

We formulate the problem of classifying fMRI-ICs as the problem of assigning the corresponding ICs-*fingerprints* to one of C classes of sources. Based on previous experience on fMRI-ICA and a preliminary analysis of data presented in this article (see results), we consider here six classes of sources (C=6): 1) The 'BOLD' class includes components that are thought to reflect consistently task-related, transiently task-related and non task-related (e.g. default state) neuronal activity; 2) the second class (MOT) includes residual motion artifacts; 3) the third class (EPI) includes the typical EPI-susceptibility artifact which is maximal in the frontal part of the brain; 4) the fourth class (VESSEL) includes physiological noise with highly localized peaks (e.g. large vessels); 5) and 6) the fifth and sixth class include noise at high spatial (SDN, spatially distributed noise) or temporal (tHFN, temporal high frequency noise) frequency. For visualization, each class is labelled with a different color (see Figure 6a for colour definition and representative exemplars of each class).

In this article we approach the classification problem using a supervised machine learning algorithm based on ls-SVMs. A complete mathematical account of this approach is beyond the scope of the article. For reader's convenience, we include a brief description of SVM and ls-SVM-based classification in a binary or multi-class case. Further mathematical details on SVM and ls-SVM can be found in Cristianini and Shawe-Taylor (2000) and Suykens *et. al.* (2002), respectively.

Support Vector Machines and least-square Support Vector Machines (binary classification).

Let us consider a training set:

$$\{\mathbf{f}_i, c_i\}; i = 1,..., I; c_i \in \{+1, -1\}; \mathbf{f}_i \in \Re^d \qquad (3)$$

where $\mathbf{f}_i$ is a d-dimensional IC-*fingerprint*, whose class c_i is known (e.g. defined by the user). Let us further suppose that the classes are linearly separable, which is equivalent to:

$$c_i(\mathbf{w}^T\mathbf{f}_i + b) \geq 0 \ \forall i, \qquad (4)$$

where **w** is the normal to an hyperplane and *b* the distance of the same hyperplane to the origin of the multidimensional space. In linear SVMs the optimal boundary between the classes is obtained by finding the hyperplane (defined by **w** and *b*) that maximizes the distance to the nearest training points of the two classes. Such points are referred to as support vectors ($\mathbf{f}_{SV}$) and lie on the marginal hyperplanes defined by:

$$c_i\,(\mathbf{w}^T\mathbf{f}_{SV} + b) = 1. \tag{5}$$

This problem can be formulated as:

$$\min_{\mathbf{w},\,b} J(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{w} \tag{6}$$

subject to:

$$c_i(\mathbf{w}^T\mathbf{f}_i + b) \geq 1, \quad i = 1,..., I. \tag{7}$$

Solution is obtained by Lagrangian methods (Cristianini and Shawe-Taylor, 2000). Classification of new IC-*fingerprints* $\mathbf{f}_{new}$ is obtained by evaluating:

$$sign\,(\mathbf{w}^T f_{new} + b). \tag{8}$$

For the more general case of non-separable classes (i.e. classes with overlapping distributions) the formulation of the SVM can be modified in order to account for misclassification errors introducing additional slack variables $\xi_i, i = 1,..., I$. The optimization problem becomes:

$$\min_{\mathbf{w},\,b,\,\xi} J(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{w} + a\,\sum_{i\,=\,1}^{I}\xi_i \tag{9}$$

subject to:

$$c_i(\mathbf{w}^T\mathbf{f}_i + b) \geq 1 - \xi_i,\ i = 1,..., I \tag{10}$$

and:

$$\xi_i \geq 0,\ \ i = 1,..., I \tag{11}$$

where *a* is a positive real constant (Suykens *et. al.*, 2002).

In the present paper, we use a variant of SVM known as ls-SVMs. In the classical SVM formulation Eq. (6-7; 9-11) the optimal boundary between different classes is obtained considering only the support vectors. In ls-SVMs each training point is weighted in order to obtain the distinguishing hyper-surface (hyperplane). The optimization problem for the general case of non separable classes is defined as:

$$\min_{\mathbf{w},\,b,\,e} J(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{w} + \gamma\,\sum_{i\,=\,1}^{I}e_i^2 \tag{12}$$

subject to:

$$c_i(\mathbf{w}^T\mathbf{f}_i + b) = 1 - e_i,\ i = 1,..., I \tag{13}$$

where γ is a positive real constant and the set of inequalities defined in Eq. (10) is replaced with a set of equalities.

SVMs (and ls-SVMs) have been modified in order to find non-linear division boundaries. In this case the data are first projected to some other Euclidean space H, using a non linear transformation, which we call φ:

$$\varphi: \Re^d \rightarrow H.$$

A linear solution is found in the space H. as in Eq. (6-7; 12-13). This corresponds to finding an optimal hypersurface in the original space such that:

$$c_i \left(\mathbf{w}^T \varphi (\mathbf{f}_i) + b \right) \geq 0 \quad \forall i. \tag{14}$$

Classification of new samples is obtained evaluating:

$$sign\,(\mathbf{w}^T \varphi (f_{new}) + b). \tag{15}$$

The use of Kernel methods (see Cristianini and Shawe-Taylor, 2000 for details) allows replacing the function φ (used for non-linear extensions) with the kernel matrix **K**. In the present paper we use the Radial Basis Function (RBF) kernel whose expression is given by:

$$K(\mathbf{f}_i, \mathbf{f}_j) = e^{-|\mathbf{f}_i - \mathbf{f}_j| / 2\sigma^2} \tag{16}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ refer to *fingerprint* values for two components i and j.

Extension to multi-class classification

As IC-*fingerprints* are to be assigned to one of six classes (i.e. $c_i \in \{1, 2, 3, 4, 5, 6\}$) we considered an extension of ls-SVMs to solve multi-class problems. The method we used is a generalization of the one-versus-all approach (for a complete review, see Allewein, 2000) and is based on Error Correcting Output Coding (ECOC) (Diettrich and Bakiri, 1995). The method requires the generation of a binary ($c \times q$) matrix (code book) that contains one row (code word) for each of the c classes and q bits. To solve the multi class problem a binary ls-SVM is trained for each of the q columns of the code book. When a new IC-*fingerprint* is tested a binary word is generated as an output of the q binary classifiers. The point is assigned to the class that corresponds to the nearest code word. ECOC is based on the concept that redundancy helps correcting errors that may be introduced along the information channel on one or more bits (for details, see Diettrich and Bakiri, 1995). Results reported in this paper are obtained using c=6 (number of IC-*fingerprint* classes) and q=5 (length of the code word).

Supervised training

For training the SVM-based classifier and optimizing the various parameters involved (e.g. σ in Eq. (16)) we employed two sets of data. A first set was generated by inspecting and separating the ICs obtained from the cb-ICA decomposition of one run (subject 1, see below) into the six classes of sources. A second set was formed with simulated samples of the same six classes. These samples were drawn from a multivariate Gaussian distribution, whose mean and standard deviation were estimated from the first set. Optimization of parameters and learning of the distinction between the six ICs classes in the multidimensional space (i.e. definition of the separating hypersurfaces) was achieved using a cross-validation technique (Mitchell, 1997), which involved iterative training on the simu-

lated data set and evaluation of performances on the real data set. We trained R=50 multi-class classifiers in order to take into account the variability introduced by the coding scheme.

Note that the use of simulated data, while introducing an assumption on the distribution of the IC-*fingerprints*, allowed keeping at a minimum the amount of real data employed during training (and thus of ICs to be inspected and labelled). An alternative approach would be to substitute the simulated data with additional real data (e.g. more data from one subject or multiple subjects).

Automatic classification of ICs

Unclassified IC-*fingerprints* were processed by each of the R classifiers. The final classification of IC-*fingerprint*s was obtained from the outcomes of the R classifications following a simple majority rule.

In order to evaluate the performances of our approach the unclassified ICs were also classified by an expert, on the basis of the visual inspection of the IC maps, time courses and power spectra (see Figure 2a-d). The ls-SVM based classifications were then compared with the expert classifications and true and false positive rates were estimated accordingly.

Furthermore, for comparison purposes, we report the classification performances obtained with a linear discriminant analysis, using the same training data and the same expert-labeled ICs as benchmark. In this analysis, covariance of the classes was assumed to be equal and was estimated by pooling the data across classes (Duda *et. al.*, 2000).

fMRI data

Visual Structure-From-Motion stimulation (block design)

We tested the proposed approach on data from a block-designed visual experiment using structure from motion (SFM) stimuli (Kriegeskorte *et. al.*, 2003). Stimuli were moving dots evoking the percept of faces or complex random shapes. The experimental conditions comprised two different types of SFM stimuli with the dots either fixed to the surface of the object (classical SFM) or moving on it (on-surface SFM). Motion and static control stimuli were also included in the stimulation protocol. In total, there were nine random-dot stimulus conditions, including moving faces (classical SFM and on-surface SFM); moving random shapes (classical SFM and on-surface SFM); static faces (on-surface SFM); static random shapes (on-surface SFM); moving-dot control matched to classical SFM; moving-dot control matched to on-surface SFM; static-dot control (for a complete description of the stimuli see Kriegeskorte *et. al.*, 2003). Each condition

appeared twice in each run, except for the two moving dot control conditions, each of which appeared only once in each run. There were, thus, 16 stimulation periods separated by 17 fixation periods. Because each period had a duration of 16 sec, an experimental run lasted 8 min. and 48 sec. The condition sequence was pseudorandom but symmetrical.

Seven subjects between 21 and 34 years of age participated in the study (average age, 25.3 years). Each of the seven subjects underwent four runs of the SFM experiment. Results presented in this article refer to the analysis of the first two runs of each subject. Subjects were instructed to continually fixate a central cross visible throughout the experiment and to classify each stimulus presented as either face or non-face as soon as they could by pressing one of two buttons (two-alternative forced choice).

Functional scans consisted of 20 transversal slices collected at 1.5 T (Magnetom Sonata; Siemens, Erlangen, Germany) using a single-shot gradient-echo echo-planar imaging sequence (in-plane resolution $3.25 \times 3.25\ mm^2$; slice thickness 5 *mm*; gap 0 *mm*; slice acquisition order interleaved; field-of-view (FOV) $200 \times 200\ mm^2$; acquisition matrix 64×64; time to repeat (TR), 2000 *msec*; time to echo (TE), 60 *msec*; flip angle (FA), 90°). Each subject underwent a high-resolution T1-weighted anatomical scan at 1.5 T (Magnetom Sonata, see above), using either a three-dimensional magnetization-prepared-rapid acquisition-gradient-echo sequence lasting 8 *min* and 34 *sec* (192 slices; slice thickness 1 *mm*; TR 2000 *msec*; TE 3.93 *msec*; FA 15°; FOV $250 \times 250\ mm^2$; matrix 256×256) or a three-dimensional T1-fast-low-angle shot sequence lasting 16 *min* and 5 *sec* (200 slices; slice thickness 1 *mm*; TR 30 *msec*; TE 5 *msec*; FA 40°; FOV $256 \times 256\ mm^2$; matrix 256×256).

The fMRI data sets were subjected to a series of pre-processing operations. (1) Slice-scan-time correction was performed by resampling the time courses with linear interpolation such that all voxels in a given volume represent the signal at the same point in time. (2) Head movements were detected and automatically corrected minimizing the sum of squares of the voxel-wise intensity differences between each volume and the first volume of the run. Each volume was then resampled in three-dimensional space according to the optimal parameters using trilinear interpolation. (3) Temporal high-pass filtering was performed to remove temporal drifts of a frequency below three cycles per run (3/528 sec). (4) After co-registration to the anatomical images collected in the same session the functional volumes were projected into Talairach space.

After these pre-processing steps, each of the 14 functional time series (7 subjects, 2 functional runs per subject) was decomposed using cb-ICA, which included PCA-based reduction of dimensionality to 60 dimensions (retaining more

than 99% of the variance/covariance of the data) and 60 x 14 independent components were extracted. Each IC was projected into the space of the eleven parameters and corresponding IC-*fingerprints* were obtained as described above. ICs from the first functional run of Subject 1 were used for expert labelling (using displays as in Figure 2) and for training the classification algorithms. The remaining components were used for testing and validating the automatic classification.

Auditory presentation of sentences (block and event-related design)

To investigate the validity of our approach in an experiment other than the SFM data set used for training, we extracted and classified - without re-training of the ls-SVM classifier - ICs from two fMRI time series collected using a different stimulation modality (auditory presentation of sentences) and different stimulation timing. The two time series are part of the publicly available Functional Imaging Analysis Contest (FIAC) 2005 data set (see Poline *et. al.*, 2006 and http://www.madic.org) and refer to the single-subject data (Subject 3). In particular, we considered the first run of the block design and the first run of the event-related design. Below an essential description of these data is provided; details on the rationale of the experiment and on acquisition and stimulation procedures can be found in Dehaene-Lambertz *et. al.* (2006, see Experiment II).

The experiment examined the functional specialization of cortical language areas using an adaptation paradigm with spoken sentences and was performed in a 3-T whole body system (Brucker, Germany). Functional images comprised 30 axial slices obtained with a T2-weighted gradient echo, EPI sequence (TR 2.5 s; TE, 35 ms; FA 80°; FOV $192 \times 192\,mm$; 64×64 pixels). Anatomical images were obtained using a high resolution ($1 \times 0.9 \times 1.4\,mm$), T1-weighted sequence. In the block design, 20 seconds blocks of six sentences were presented in which either the speaker voice or the linguistic content of the sentences, or both, were repeated. Stimulation blocks were followed by 9 seconds 'silence' blocks. In the event-related design one sentence was presented every 3333 ms. The same conditions as in the block design were presented, but they were defined by the transition between two sentences (Dehaene-Lambertz *et. al.*, 2006).

Following standard preprocessing (see Goebel *et. al.*, 2006 for details), we decomposed the data sets using cortex-based ICA, which included PCA-based dimensionality reduction to 60 dimensions, and characterized extracted components using IC-*fingerprints*. We then proceeded in classifying the IC-*fingerprints* with the ls-SVM based classifier, which had been trained on the visual SFM experiment as described above.

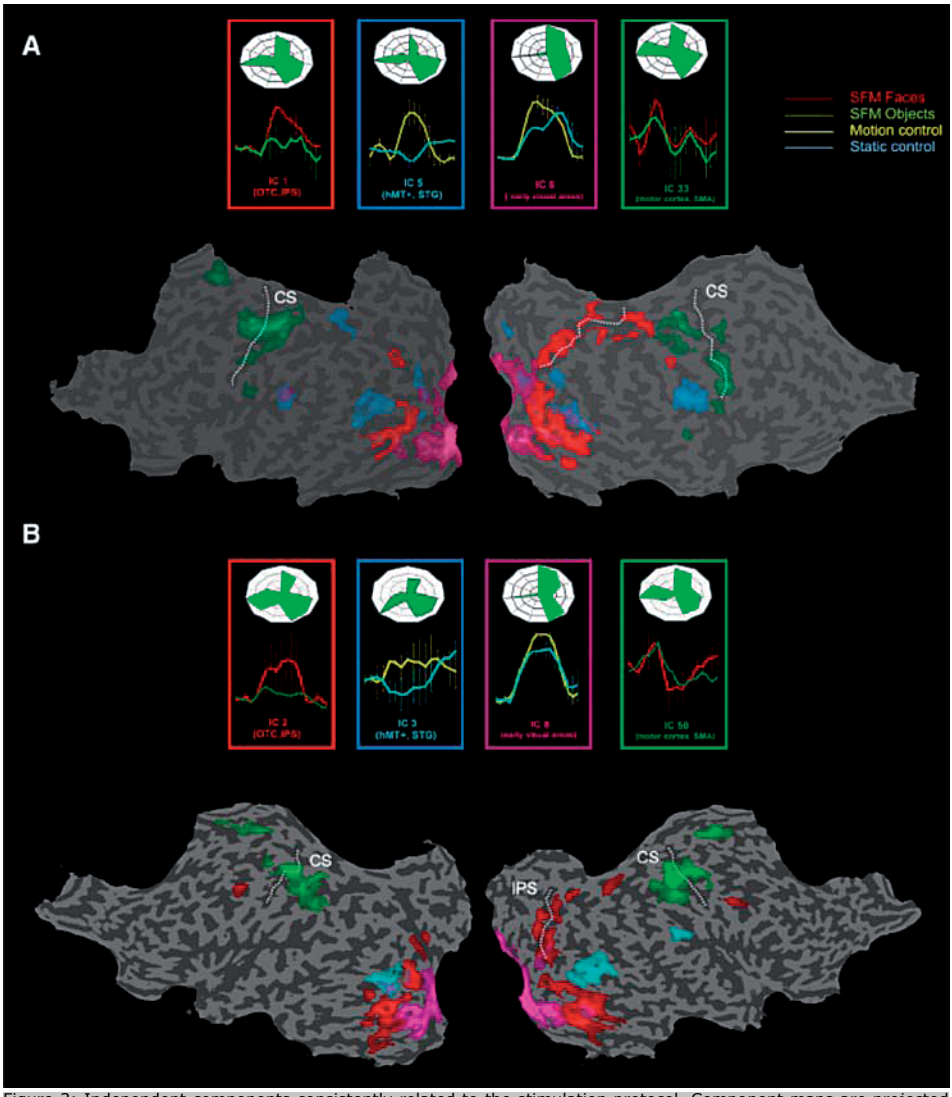

Figure 3: Independent components consistently related to the stimulation protocol. Component maps are projected on the flattened representation of the subject's cortex: a) Subject 1 (BS), run 1 (training set); b) Subject 2 (AH), run 1 (automatic classification).
Colored-matched inserts show condition-based averages of the IC time courses for the most relevant comparisons. Numbering of ICs is based on the order of extraction in the ICA decomposition. CS = central sulcus; IPS = intraparietal sulcus; STG = superior temporal gyrus; SMA = supplementary motor area; OTC = occipito-temporal cortex.

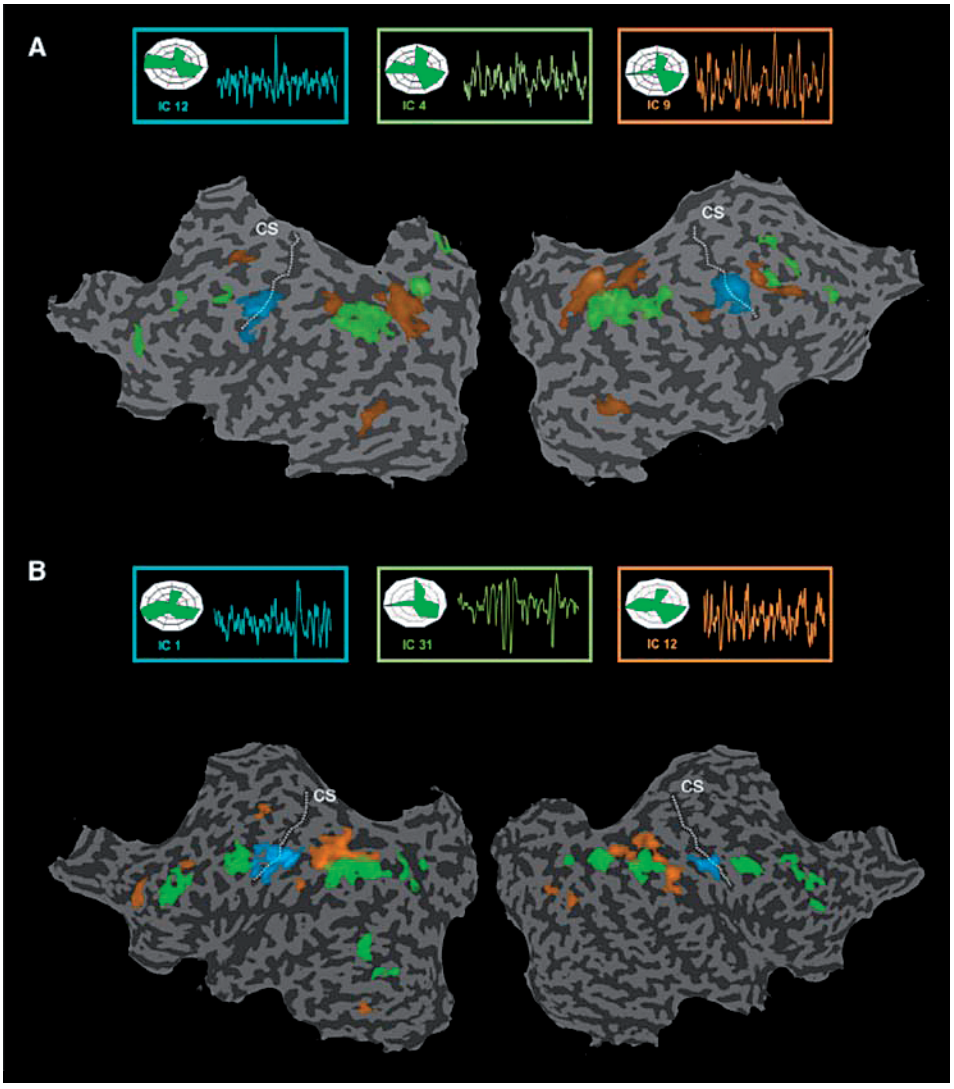

Figure 4: Other independent components reflecting neurophysiological sources. Component maps are projected on the flattened representation of the subject's cortex; a) Subject 1 (BS), run 1 (training set); b) Subject 2 (AH), run1 (automatic classification).
Colored-matched inserts show the IC time courses. Numbering of ICs is based on the order of extraction in the ICA decomposition.

Results

ICA analysis and characterization of the ICs

Figure 3a, 4a illustrate the results of the ICA decomposition in Subject 1 (run1) in the visual SFM experiment. Together with IC maps, visualized on the flattened representation of the subject's cortex, and time courses these figures include the eleven-dimensional *fingerprint* of each IC.

As expected, a subset of components included ICs that were consistently related to the stimulation protocol (Figure 3a). The spatial maps of these components encompassed a widespread set of visual ventral and parietal areas, whose activation has been detected with univariate hypothesis-driven analysis (Kriegeskorte *et. al.*, 2003). In particular, ICA decomposition highlighted three distinct spatiotemporal patterns reflecting: 1) activation in early visual areas (violet component in Figure 3a); 2) co-activation of ventral visual regions including the lateral occipital cortex (LOC) and the fusiform-face area (FFA) and regions along the intra-parietal sulcus (IPS) (red component in Figure 3a) and 3) co-activation in the motion complex hMT+ and temporo-parietal regions (light blue in Figure 3a). Condition-based averages of these component time courses reflect the different functional role of these networks in the processing of the SFM stimuli (see insert in Figure 3a) and highlight a strong selectivity for SFM faces compared to SFM control surfaces in the case of the ventro-parietal component (see Discussion). A fourth task-related component (green in Figure 3a) reflected the activation in the hand region of the central sulcus (CS) during the motor response at the beginning of each block. Importantly, IC-*fingerprints* associated to these components showed a high degree of similarity, with a high value of degree-of-clustering and temporal autocorrelation. Note also the high contribution of low and mid frequency range in all these components, with a sharper peak for the primary visual component in the range that includes the stimulation frequency.

Inspection of the ICs with *fingerprints* most similar to those of stimulus-related ICs revealed interesting activation patterns in other specific regions or in functionally-connected networks. Although the time courses of these ICs are not (or not consistently) related to the stimulation protocol, their spatial layout together with the statistical properties of their histogram and their spectral properties are very similar to those of consistently task-related ICs, suggesting a common neuronal/BOLD nature of the underlying sources. Figure 4a shows three of such ICs, including two segregated fronto-parietal networks (green and brown) and a component located in the lower part of the motor cortex bilaterally (light blue).

Other ICs with comparable *fingerprints* reflected the typical *default-mode* networks (maps not shown), consistently found in several recent works (Raichle *et. al.*, 2001, Greicius *et. al.*, 2003).

The remaining components reflected the contribution of artifactual sources, as it is commonly found in fMRI-ICA decompositions (McKeown *et. al.*, 2003, Thomas *et. al.*, 2001). These sources included large vessels, subject's motion, signal changes due to EPI-susceptibility artifacts and noise at high spatial and high temporal frequencies. The ICs reflecting the same type of sources were associated with similar and distinctive *fingerprints*. Typical samples of these ICs and corresponding *fingerprints* are shown in Figure 6a.

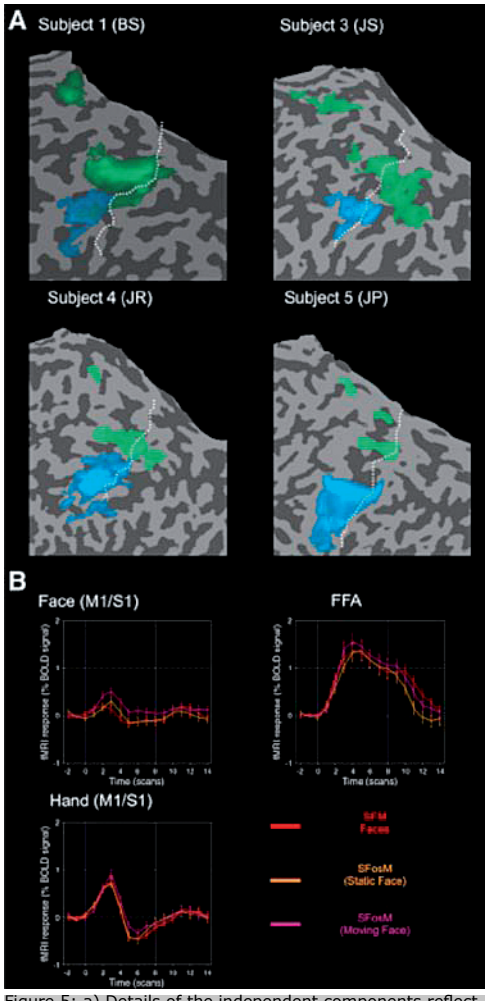

Figure 5: a) Details of the independent components reflecting the response in the 'face' (blue) and 'hand' (green) sensorimotor cortex along the central sulcus. Components are projected on the flattened representation of the left hemisphere of four different subjects; b) average response to stimulation conditions involving the perception of faces in three distinctive regions.

Automatic classification of the ICs

Visual structure-from-motion stimulation

The analysis of Subject 1 (run 1) described above served as the basis for generating a training data set. After training and optimization of the ls-SVM classifier, ICs resulting from the ICA-decomposition of remaining data sets (13 runs, run 2 of subject 1, runs 1 and 2 of subjects 2-7) were submitted to the classifier and automatically labelled.

In each subject, the automatic classification identified as 'BOLD' a number of ICs ranging from a minimum of eight to a maximum of fifteen. In all cases, the class of BOLD ICs included the subset of consistently task-related components, accounting for the activation of early visual areas, hMT+, ventral and parietal areas, and 'hand' sensorimotor cortex (see e.g. subject 2 in Figure 3b). The *default*

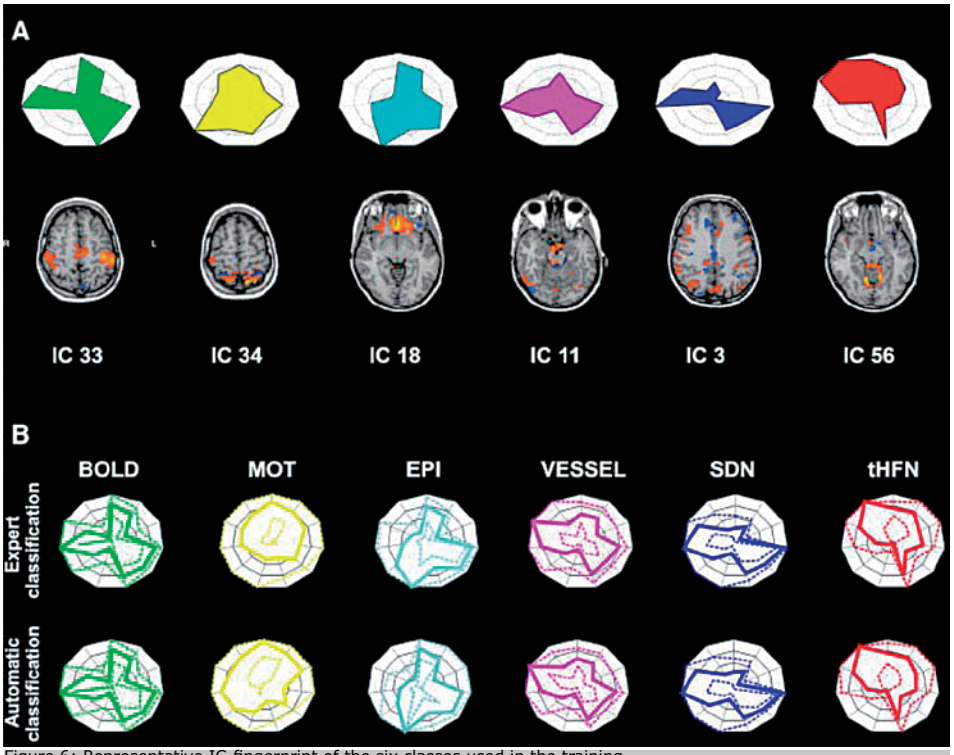

Figure 6: Representative IC-fingerprint of the six classes used in the training.
a) Fingerprints and transversal slice of each class-representative IC in the training data set. Numbering of ICs is based on the order of extraction in the ICA decomposition.
b) Median (bold line), 5% and 95% quantiles (dashed lines) IC-fingerprints, for each class as obtained in the expert (top row) and automatic (bottom row) classification.

mode ICs were also consistently included in the BOLD class. Furthermore, in all subjects this same class included the fronto-parietal ICs and the IC with localized activity along the CS (see e.g. subject 2 in Figure 4b). To further investigate this latter IC we analysed in detail its spatial layout and the time course of activity around the component peaks (Figure 5). Consistently across subjects, the prominent peak of this IC was located in correspondence of the 'face' region of the primary sensorimotor cortex, located along the CS inferiorly and medially to the 'hand' sensorimotor region (Lotze *et. al.,* 2000). Condition-based averaging, across all subjects, of the time course of activation in the individually ICA-defined 'face' region revealed a transient activation at the stimulus onset in all conditions involving the perception of faces. Activation time courses in the FFA and in the 'hand' M1/S1 are reported for comparison.

The ls-SVM based algorithm classified accurately most of the artifacts-related ICs in the data set. The EPI susceptibility artifact was successfully labelled in eight out of thirteen runs. The spatio-temporal features for this class were very

similar to those observed for the class 'BOLD' (high degree of clustering and high autocorrelation at the first lag) but a much higher power contribution at the lowest frequency band was present. The SDN class was consistently characterized by a low degree of clustering while resembling the class BOLD in the other parameters. The highest degree of similarity was observed between the class VESSEL and the class BOLD. The class tHFN was consistently represented across runs by a very characteristic *fingerprint* (high spectral contribution in the highest frequencies, and low kurtosis values). The class that presented the most variable *fingerprints* was the class reflecting residual motion artifacts (MOT).

Figures 6b and Table 1 detail the analysis of the performances of the ls-SVM based classification. We evaluated the correspondence of the classification obtained using the SVM based classifier with the expert classification of all ICs.

Figure 6b shows, for each class, a representative *fingerprint* obtained by connecting the median values along each dimension (bold line). In order to highlight the within-class range of variability, two additional *fingerprints* are superimposed to each class-representative *fingerprint*. These *fingerprints* are obtained by connecting the 5% and 95% quantiles along each dimension (dashed lines). The bottom row shows the results obtained from the ls-SVM based classifier. For comparison, the top row shows the representative *fingerprints* of the classes obtained by expert labelling of ICs, uniquely based on the inspection of IC time courses and maps. The high correspondence between the top and bottom row *fingerprints* for the classes BOLD, SDN, tHFN indicates that the ls-SVM classifier operates as a human expert for these classes. More discrepancy and variability is noticeable for the classes VESSEL, EPI and MOT.

Table 1
Performances of the ls-SVM based classifier on the classification of IC-fingerprints when compared with the expert classification

	BOLD		MOT	EPI	VESSEL	SDN	tHFN
	task related	others					
true positive	100% (100%)	90% (84%)	35% (48%)	61% (82%)	61% (60%)	100% (83%)	100% (74%)
false positive	0% (0%)	4% (5%)	0% (5%)	2% (3%)	2% (10%)	4% (3%)	3% (1%)

True positive rate is defined as the ratio between the number of components correctly assigned to a class and the total number of components in that class.
False positive rate is defined as the ratio between the components incorrectly assigned to a class and the total number of components not belonging to that class.
For comparison, performances of a linear discriminant classifier are reported in parenthesis.

A characterization of the performances in terms of true positive and false positive rates is presented in Table 1. The automatic classifier reached a 94% true positive rate for the class BOLD, signifying that in 94 % of the cases the classification algorithm and the human expert identically labelled these components. Restricting the comparison to the task related ICs (i.e. to the ICs with a clear interpretation) the overlap between expert and automatic classification increased to 100%. Similarly, a 100% true positive rate was achieved for the classification of ICs in classes SDN and tHFN. For the classes EPI and VESSEL, the true positive rate was 61%. Performances dropped for the class MOT (35% true positive rate). The false positive rate, computed as the ratio between the ICs incorrectly assigned to a class and the total number of components not belonging to that class. For each class this rate was below 5%.

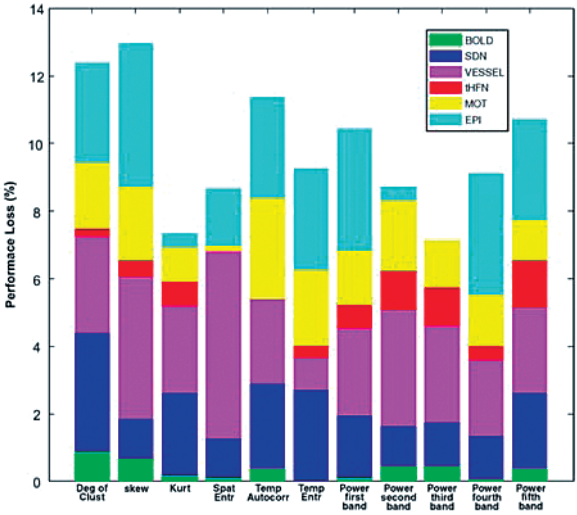

Figure 7: Percent loss of performance (in terms of correspondence with the expert classification) of the automatic classification after removal of each of the parameters forming the fingerprint. Loss values are detailed for all six classes and are referred to the performance achieved by the original classifier.

Table 1 also reports the classification performance obtained using a linear discriminant analysis. When compared to the expert classification, also this linear classification algorithm showed a good correspondence. However, it can be noticed that the linear discriminant classifier performed slightly worse than the ls-SVM classifier, except for MOTION and EPI classes.

We further assessed the informative nature of all measures that we included in the *fingerprints* by re-evaluating the performances of the automatic classification after removal of each measure. In all cases, removal of individual measures decreased the classification performances. Figure 7 reports for each of the six classes the percent drop of performances of the reduced classifiers with respect to the ls-SVM based classification obtained using all measures.

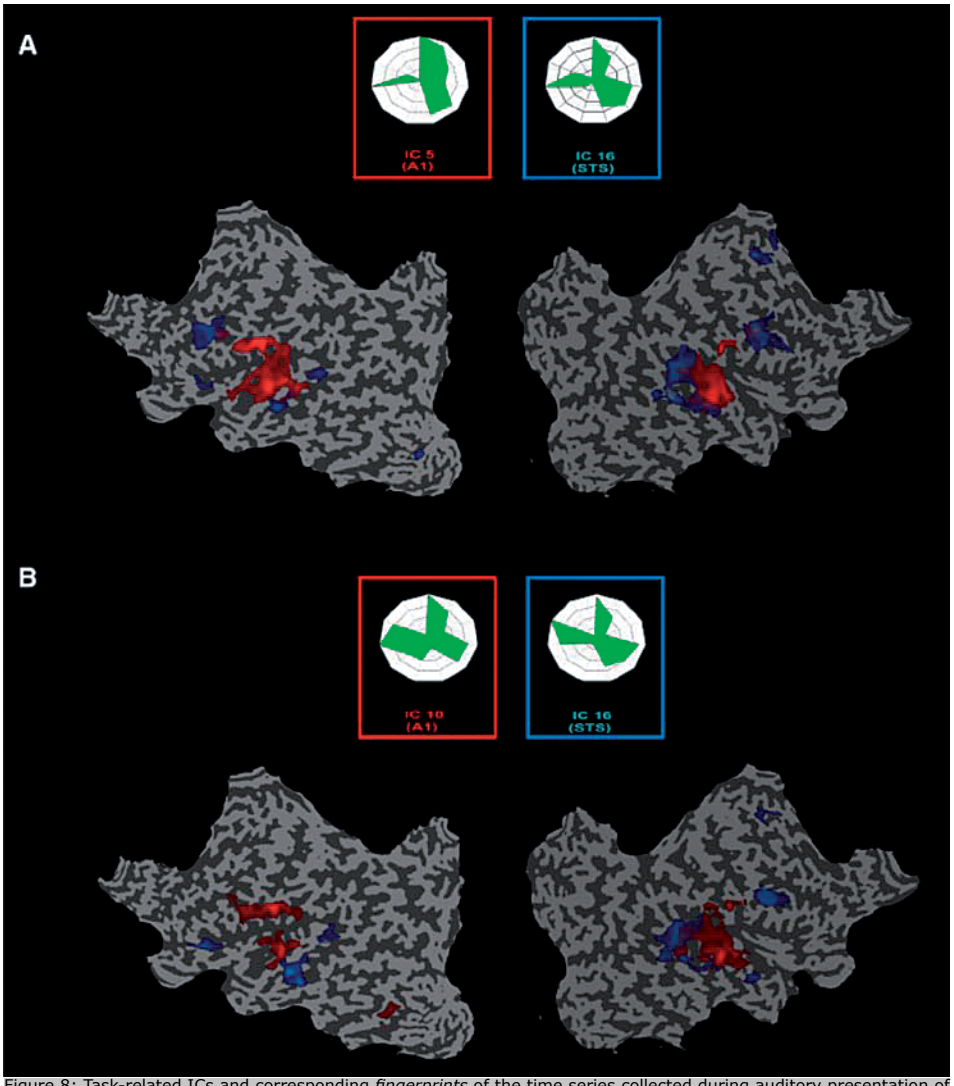

Figure 8: Task-related ICs and corresponding *fingerprints* of the time series collected during auditory presentation of sentences. ICs were automatically identified by the ls-SVM classifier trained on the SFM data. a) Block design experiment. b) Event related experiment. Red: auditory component; blue: temporo-frontal component.

Auditory presentation of sentences

We verified the capability of our approach in identifying 'meaningful' components in a new experiment by extracting and classifying - without re-training of the ls-SVM classifier - ICs from the two fMRI time series of the FIAC data set.

Figure 8 illustrates the spatial layout and IC-*fingerprint*s of two of the components that the classifier identified as 'BOLD' for the block (Figure 8a) and for the event-related (Figure 8b) time-series. The first component (red colour) reflects

the stimulus-related activation evoked by the auditory presentation in bilateral primary and secondary auditory cortical regions. The second component (blue colour), which includes a more distributed temporo-frontal network with clusters located along the superior temporal sulcus and gyrus (STS/STG) and the inferior frontal gyrus, presumably reflects the linguistic processing of the sentences (see also Dehaene-Lambertz *et. al.*, 2006 and Goebel *et. al.,* 2006). Note the visual resemblance of the IC-*fingerprints* obtained in these latter time series with those obtained in the SFM data set, despite the considerable differences between the two experiments. As in the SFM experiment, other ICs classified as BOLD were the default mode ICs (not shown, see Goebel *et. al.* 2006). In these classifications the true positive rate (i.e. the correspondence with the expert classification) was on average 95% for the class BOLD (block: 100%, event-related: 90%), 100 % for the class tHFN, 81% for the class SDN (75%, 88%), 37% for the class VESSEL (40%, 33%). No component was classified as MOT or EPI.

Discussion

This study illustrates a general approach for the characterization and classification of fMRI independent components. Differently from conventional univariate statistical analyses, in which a small set of predefined hypotheses is tested, spatial maps (and associated time courses) obtained in fMRI-ICA are determined solely by the intrinsic structure of the data. Such a data-driven analysis provides an attractive opportunity for a blind detection of potentially interesting spatio-temporal patterns (such as networks of functionally connected brain regions) and structured artifacts. At the same time, the application of ICA to fMRI data analysis challenges the experimenter with the problem of selecting a 'meaningful' subset from the large set of obtained components. Whereas this problem ultimately requires interpretation - grounded in the knowledge of an expert - in the first part of this article we show that certain computable properties of the components can be utilized to guide the exploration of the fMRI-ICA results. In the second part of the article, we show that expert knowledge on fMRI-ICs can be transferred to a machine learning algorithm, allowing an accurate and fast automatic classification of 'signal' and 'noise' components.

In the analysis of the data used for training, the IC-*fingerprint* proved to be a powerful tool for the inspection of the ICs, independently of the automatic classification, Neurophysiologically plausible (BOLD) ICs were characterized by a distinctive *fingerprint* resulting from a high spatial and temporal structure and a prominent power contribution in the 0.01-0.1 Hz frequency range. This was not

surprising in the case of consistently task-related ICs because of their simple relation to the visual stimulation (0.03 Hz) employed in the presented experiments. However, this also held true for other BOLD components which did not correlate strongly with the stimulus, such as the fronto-parietal components, the 'face' motor component and the *default mode* components, which reflected interesting effects not detected in the conventional GLM analysis (Kriegeskorte *et. al.*, 2003). These observations are in line with recent work suggesting that intrinsic neurovascolar oscillations in functionally connected regions are reflected in this frequency range (Cordes *et. al.*, 2001, Raichle *et. al.*, 2001, van de Ven *et. al.*, 2004, Greicius *et. al.*, 2003). Similarly, components reflecting the same type of sources resulted in *fingerprints* with characteristic and recognizable shapes, indicating that visualizing fMRI-ICs using this display is a viable approach for their grouping.

IC-*fingerprints* were formed using a set of eleven measures (dimensions). These measures were chosen on the basis of previous experience (Formisano *et. al.*, 2002) and of theoretical and heuristic considerations on global statistical and spatio-temporal properties of the fMRI signal (see Appendix 1). In the fMRI data sets we analysed, the selected set of measures proved to be informative with respect to the problem of differentiating the various source types. The informative nature of the measures forming the IC-*fingerprints* is supported not only by a qualitative inspection of the *fingerprint* shapes but also by the results of the re-evaluation of the classification performances after removal of each of the measures. For each individual measure, the observed drop of performances can be taken as an indication of its relevance to the classification problem.

While showing that the IC-*fingerprint* representation is of practical relevance and utility in the analysis of fMRI data, our work does not allow concluding that the proposed representation is 'optimal' in the sense of classification performance. IC-*fingerprints* can be easily extended (and possibly improved) so as to incorporate a larger set of general features (such as stationary index of the time-series or spatial smoothness of the maps). The inclusion of a new (set of) measures requires a new training of the classifier and a new analysis of the classification performance.

The ls-SVM algorithm learned expert knowledge on ICs from a very small training data set and was able to generalize this knowledge across runs and subjects, even with different order of stimulation conditions. The algorithm achieved very high levels of correspondence with the expert in detecting task-related and other BOLD ICs in general. In each subject, the ICs labelled as 'BOLD' included expected (i.e. consistently stimulus-related) responses in early visual areas, hMT+ and temporal regions, ventral occipito-temporal and inferior parietal

regions. Interestingly, this ventro-parietal component indicates co-modulation of activity in regions of the 'what' and 'where' processing streams during the encoding of complex surfaces. Unexpectedly, but with high intra- and inter-subject reproducibility, the 'BOLD' class also included two separate ICs consisting of two spatially segregated bilateral fronto-parietal networks and one component consisting of the face region of the sensorimotor cortex and adjacent precentral sulcus (see Figure 4). Although a conclusive interpretation of these findings is not possible without confirmatory experiments, it is interesting to relate these observations with results from recent neuroscientific literature.

The systematic segregation of the clusters in the two fronto-parietal components with a high power contribution in the low-frequency bands suggest that these set of regions belong to functionally connected networks similar to those previously reported in studies of resting state with fMRI (Cordes *et. al.*, 2001, van de Ven *et. al.*, 2004, Greicius *et. al.*, 2003, Raichle *et. al.*, 2001), electroencephalography (Tucker *et. al.*, 1986) and direct neuronal recording (Leopold *et. al.*, 2003). In the present case, however, subjects were involved in a perceptual task and thus it cannot be excluded that these ICs reflect neuronal activity with a more complex relation to the stimulus (e.g. control or suppression of eye-movements, changes in attentional state of the subjects) and which is not captured by the linear relation of the IC time courses with the stimulation protocol. The interpretation of these ICs may be aided, in future studies, by simultaneous collection of additional data (e.g. recording of eye movement data and electroencephalographic activity).

The consistent presence in all subjects of a component with high values clustered around the 'face' sensorimotor regions suggests a transient involvement of these regions in the SFM perceptual task. At the beginning of each stimulation block, subjects were asked to indicate, by pressing one of two buttons, whether the stimulus presented was either a face or a non-face control, as soon as enough evidence was extracted from the moving dot stimuli. It is plausible that subjects solved such a task by reverting to the use of implicit motor (or somatosensory) imagery (Parsons *et. al.,* 1995), i.e. they imagined their face moving (or they imagined dots moving on their face) in order to recognize whether the SFM stimulus was a moving face or not. Alternatively, this effect may be interpreted as an automatic response, generating the interesting hypothesis that, in addition to or as part of the mirror system, somatotopic regions of the primary motor and somatosensory cortex are involved in the recognition of a moving body part (Rizzolatti *et. al.,* 2004). To test these hypotheses and, more in general, to elucidate the role of primary sensorimotor regions in the perception, recognition and processing of faces, we are currently designing fMRI experiments that exploit

parametric manipulations of SFM stimuli.

This transient activation in the face-motor region was undetected in a conventional GLM analysis, in which stimulus-related BOLD responses were modelled as sustained responses. In general, the use of more flexible models (e.g. Fourier basis functions, FIR models) or the explicit inclusion of a transient predictor may also have lead to a similar result. However, relevant differences between ICA and GLM-based approaches remain and are to be noted. First, ICA sensitivity is not influenced by the trial-by-trial variability of its time-course, which is (conventionally) not modelled in GLM-based approaches (Duann *et. al.*, 2002). Second, depending on the correspondence between hemodynamics and models employed, GLM sensitivity may be different in different regions of the brain. Third, the detection of weak, transient effects with ICA may be favoured by the 'automatic' separation from confounds and noise sources, which are not known a priori and thus would not be included in the GLM model as confounds.

Automatic classification was less accurate for some of the artifactual classes, especially in the case of residual motion signal effects. The lower performances for these classes are most likely due to the small number of samples employed in the training. An additional explanation is that motion-related source processes, while being systematically extracted by spatial ICA because of their spatial structure (Kochiyama *et. al.,* 2005), are much less stationary in the temporal domain (see Esposito *et. al.,* 2003). This non-stationary behaviour may reduce the effectiveness of the temporal measures and of the proposed representation in general. Especially in these cases, the performances of the classifier may be enhanced by adding new parameters that take these aspects properly into accounts, the dimensionality of the IC-*fingerprints* not being a relevant problem.

The classifier trained on the SFM data was able to successfully select 'meaningful' components in the block and event-related FIAC data set. Notably, these successful selections were obtained without further training and on time series collected using a different MR scanner and magnetic field strength (3 vs 1.5 Tesla), a different TR (2.5 s vs 2.0 s) and different stimulation timing and modality (auditory vs visual) compared to the data used for training. This result suggests that the proposed approach can be directly applied to a great variety of datasets. More generally, however, it should be considered that larger differences in acquisition parameters (especially TR), in the stimulation timing and/or pre-processing choices (e.g. degree of spatial/temporal smoothing) may alter significantly the spatiotemporal properties of the fMRI time series, thus affecting significantly the shape of the IC-*fingerprint* and, consequently, the performance of the classifier. In such cases, it is necessary to perform a new training of the classifier using the strategy described for our training data set. Additional empirical work is ultimately

required to examine to what extent and under which circumstances this extra training is required.

When limited to the detection of 'BOLD' ICs, our approach leads to a major reduction of the number of components to be inspected and interpreted. A reduction of the number of ICs can be also achieved by focusing on ICs that are most common across the sample subjects, as is the case in recently proposed group-ICA approaches. An important difference, however, is that these methods rely on spatial (Calhoun *et. al.*, 2001, Esposito *et. al.*, 2005), temporal (Svensen *et. al.*, 2002) or spatiotemporal (Beckmann *et. al.*, 2005) consistency across subjects and do not make an explicit distinction between common 'signal' (e.g. BOLD) and common 'noise' (e.g. EPI-susceptibility artifact). Conversely, the reduction achieved with the proposed method is at the level of the single-subject and allows detection of subject-specific ICs independently of their spatiotemporal matching with other subjects and independently of their contribution to the variance-covariance of the group data set. This may be a very favourable option in, e.g., clinical cases (see van de Ven *et. al.*, 2005). In multi-subject studies, the proposed approach can be usefully combined with a recently proposed algorithm for grouping ICA components across runs and subjects (Esposito *et. al.,* 2005). In fact, a first-level reduction of the whole set of estimated components to the class of neurophysiologically 'meaningful' components can be used to increase the power of the subsequent grouping by reducing the cross-subject interference of signal and noise components.

A final methodological consideration concerns the fact that IC-*fingerprints* are derived, in the present implementation, by post-hoc measures of the fMRI-ICs. It may be interesting, in future developments, to tailor fMRI-ICA algorithms by including such global expectations on the temporal, spatial and distributional properties of the 'meaningful' components directly in the principle of estimation (Calhoun *et. al.*, 2005) or in the contrast function (Valente *et. al.,* 2005). This may allow increasing the sensitivity of current ICA algorithms in detecting spatiotemporally structured fMRI sources and sorting the interesting components during the extraction.

Conclusions

A technique for the automatic classification and the selection of relevant ICA components in fMRI data has been presented and validated. Its most important feature is that it matches the hallmark of ICA, i.e. blind detection of unexpected, yet plausible and interesting, neural (BOLD) activation patterns. The proposed solution facilitates the use of ICA for the explorative analysis of complex fMRI data sets. In combination with an appropriate choice of specific measures and heuristics, a similar approach to the selection of the components could be extended to other applications of ICA, such as the analysis of electro- and magnetoencephalography data.

Acknowledgments

The authors are grateful to Nikolaus Kriegeskorte and Bettina Sorger for kindly providing the experimental data and for insightful discussions, and to Jean-Baptist Poline and the MADIC's Team (Orsay, France) for making the FIAC data available. Financial support from NWO (MaGW-VIDI grant 452-04-330) to EF is gratefully acknowledged.

Appendix 1

Measures derived from IC map values

Components extracted using commonly-employed ICA algorithms exhibit a non-Gaussian (typically super-Gaussian) distribution of voxel values. The inclusion in the IC-*fingerprint*s of measures that estimate sparseness, asymmetry and information content (kurtosis, skewness, entropy) of the IC distributions was suggested by previous related work (Formisano *et. al.*, 2002; Suzuki *et. al.*, 2001) and by the empirical observation that sources of the same type may exhibit similar distributions of map values. As these measures do not convey any information regarding the spatial structure of obtained maps (i.e. the spatial proximity of voxel values is ignored), we also included a measure (degree of clustering) that exploits the fact that meaningful processes tend to have a well-defined spatial structure (Formisano *et. al.*, 2002).

Kurtosis is a measure of the sparseness of a distribution; it is zero for Gaussian distributions (Suhir, 1997). For each IC, *kurtosis* was estimated as:

$$kurt_i = \frac{\sum_{k=1}^{N} z_{i,k}^4}{N} - 3$$

where $z_{i,k}$ represents the value of the k-th voxel of the i-th component and N is the number of voxels. The normalized vector entering the *fingerprints* was obtained by linear scaling transform of |ln (**kurt**)|.

Skewness is a measure of the asymmetry of the distribution; it is zero for Gaussian distributions (Suhir, 1997) and has been used for tailoring ICA decompositions in fMRI (Suzuki *et. al.,* 2001). For each IC (after mean removal and variance normalization), skewness was estimated as:

$$skew_i = \frac{\sum_{k=1}^{N} z_{i,k}^3}{N}$$

where $z_{i,k}$ represents the value of the k-th voxel of the i-th component and N is the number of voxels. The normalized vector entering the *fingerprints* was obtained by linear scaling transform of |ln (**skew**)|.

Spatial entropy is a measure of the information content of a spatial distribution. Information content (and thus spatial entropy) is expected to be higher for components with widely distributed values compared to components with a nar-

row distribution. For each IC, spatial entropy was estimated as:

$$H_i = \sum_{b=1}^{N_b} hs_i(b) \cdot \log_2(hs_i(b))$$

where hs_i represents voxel values histogram of the i-th component computed over N_b bins. The normalized vector was obtained by linear scaling transform of $|\ln(\mathbf{H})|$.

Degree of clustering is a measure of the spatial structure of the component. For each IC, the number of voxels (N_{tot}) exceeding a threshold value in the z-normalized map ($|z|>2.5$, see Methods) and the size of the subset of these voxels (N_{clu}) belonging to a 3D cluster of minimum extension (270 mm^3) were computed. The degree of clustering was then defined as $CLU_i = N_{clu} / N_{tot}$ with values in the interval [0, 1].

Measures derived from IC time course and frequency spectrum.

The inclusion of measures of temporal structure (one-lag autocorrelation, temporal entropy) in the IC-*fingerprint*s and spectral decomposition was motivated by the following considerations. In spatial ICA no explicit constraint is posed on the time course of the sources. Nevertheless, due to the nature of the hemodynamic response, components reflecting neurophysiologically interesting sources are expected to present higher temporal structure than components reflecting noise and a concentration of spectral power at the low frequencies (0.01-0.1 Hz, Cordes *et. al.*, 2001; van de Ven *et. al.*, 2004). Conversely, frequencies above 0.1 Hz can reflect the effect of aliasing of cardiac and respiration artifacts while the typical susceptibility artifact due to the EPI sequence presents an effect at very low frequencies below 0.01 Hz.

One-lag serial autocorrelation is a measure of temporal structure. A "smooth" signal such as the BOLD responses is expected to have high values of autocorrelation (Baumgartner *et. al.*, 2000). White noise conversely is characterized by an autocorrelation function that is one at zero lag and zero everywhere else. For each IC, one-lag autocorrelation was estimated as:

$$r_i = \frac{\frac{1}{T-1}\sum_{t=1}^{T-1} a_i(t) \cdot a_i(t+1)}{\frac{1}{T}\sum_{t=1}^{T} a_i(t)^2}$$

where a_i is the time course of the i-th component and T is the number of time

points. The normalized vector was obtained considering $|\mathbf{r}|$.

Temporal entropy is a measure of information content of the time course of a component. White noise is expected to have higher temporal entropy than time structured (periodic) signals. For each IC, temporal entropy was estimated as:

$$H_i = \sum_{b=1}^{N_b} ht_i(b) \cdot \log_2(ht_i(b))$$

where ht_i represents time course values histogram of the i-th component computed over N_b bins. The normalized vector was obtained by linear scaling transform of $\exp(|\mathbf{H}|)$.

Power contribution The power spectrum density for each IC time course was computed using Welch's method (see Childers, 1997). Five measures were obtained by considering the relative contribution of 5 different frequency bands ([0, 0.008 Hz], [0.008, 0.02 Hz], [0.02-0.05 Hz], [0.05-0.1 Hz] and [0.1-0.25 Hz]) to the overall spectral power. The matrix containing the spectral measures of all the ICs sorted by rows was normalized by scaling each row to its maximum value and then scaling each column to its maximum value.

References

Allwein, E., Schapire, R., Singer, Y. 2000. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. Journal of Machine Learning Research 1, 113-141

Bartels, A., Zeki, S., 2005. Brain dynamics during natural viewing conditions: a new guide for mapping connectivity in vivo. Neuroimage 24(2), 339-49.

Baumgartner, R., Somorjai, R., Summers, R., Richter, W., Ryner, L., 2000. Novelty indices: identifiers of potentially interesting time-courses in functional MRI data. Magnetic Resonance Imaging 18(7), 845-850.

Beckmann, C.F., and Smith, S.M., 2005. Tensorial extensions of independent component analysis for multisubject FMRI. NeuroImage 25(1), 294-311.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Hum Brain Mapp. 13(1), 43-53.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Mapp. 14(3), 140-51.

Calhoun, V.D., Adali, T., Stevens, M.C., Kiehl, K.A., and Pekar J.J., 2005. Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ICA analysis. NeuroImage 25(2), 527-538.

Castelo-Branco, M., Formisano, E., Backes, W., Zanella, F., Neuenschwander, S., and Singer, W., 2002. Activity patterns in human motion-sensitive areas depend on the interpretation of global motion. Proc. Natl. Acad. Sci. U. S. A. 99, 13914–13919.

Childers, D.G. 1978. Modern spectral analysis. New York: IEEE Press.

Cristianini and Shawe-Taylor 2000. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.

Cordes, D., Haughton, V.M., Arfanakis, K., Carew, J.D., Turski, P.A., Moritz, C.H., Quigley, M.A., Meyerand, M.E., 2001. Frequencies contributing to functional connectivity in the cerebral cortex in „resting-state“ data. Am J. Neuroradiol. 22(7), 1326-33.

Dehaene-Lambertz, G., Dehaene, S., Anton, J., Campagne, A., Ciuciu, P., Dehaene, G.P., Denghien, I., Jobert, A., LeBihan, D., Sigman, M., Pallier, C., Poline, J., 2006. Functional segregation of cortical language areas by sentence repetition. Hum. Brain Mapp. 27(5), 360-371.

Diettrich, T., Bakiri, G., 1995. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263-286.

Duann, J-R., Jung, T-P., Kuo, W-J., Yeh, T-C., Makeig, S., Hsieh, J-C., Sejnowski, T.J., 2002. Measuring the variability of event-related BOLD signals. NeuroImage 15, 823-25.

Duda, R.O., Hart, P.E., Stork, D.G., (2000). Pattern Classification (2nd Edition). Wiley.

Esposito, F., Seifritz, E., Formisano, E., Morrone, R., Scarabino, T., Tedeschi, G., Cirillo, S., Goebel, R., Di Salle, F., 2003. Real-time independent component analysis of fMRI time series. NeuroImage 20(4), 2209-24.

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., and Di Salle, F., 2005. Independent component analysis of fMRI group studies by self-organizing clustering. NeuroImage 25(1), 193-205.

Formisano, E., Esposito, F., Kriegeskorte, N., Tedeschi, G., Di Salle, F., and Goebel, R., 2002. Spatial independent component analysis of functional magnetic resonance imaging time series: characterization of the cortical components. Neurocomputing 49(1–4), 241–254.

Formisano, E., Esposito, F., Di Salle, F., Goebel, R., 2004. Cortex-based independent component analysis of fMRI time series. Magn Reson Imaging 22(10), 1493-504.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum. Brain Mapp. 27(5), 392-401.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc. Natl. Acad. Sci. U. S. A. 100, 253-258.

Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. Proc. Natl. Acad. Sci. U. S. A. 101, 4637-4642.

Hyvärinen, A., 1999. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10(3), 626-

634.

Kochiyama, T., Morita, T., Okada, T., Yonekura, Y., Matsumura, M., Sadato, N., 2005. Removing the effects of task-related motion using independent-component analysis. Neuroimage 25(3), 802-14.

Kriegeskorte, N., Sorger, B., Naumer, M., Schwarzbach, J., van den Boogert, E., Hussy, W., Goebel, R., 2003. Human cortical object recognition from a visual motion flowfield. J Neurosci. 23(4), 1451-63.

Leopold, DA, Murayama, Y, Logothetis, NK., 2003. Very slow activity fluctuations in monkey visual cortex: implications for functional brain imaging. Cereb Cortex 13, 422– 433.

Lotze, M., Erb, M., Flor, H., Huelsmann, E., Godde, B., and Grodd, W., 2000. fMRI Evaluation of Somatotopic Representation in Human Primary Motor Cortex. NeuroImage 11(5), 473-481.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp. 6(3), 160-88.

McKeown, M.J., Hansen, L.K., Sejnowski, T.J., 2003. Independent component analysis of functional MRI: What is signal and what is noise? Current Opinion in Neurobiology 13(5), 620-629.

Mitchell, T. M. 1997. Machine Learning. McGraw Hill International Editions.

Moritz, C.H., Rogers, B.P., Meyerand, M.E., 2003. Power spectrum ranked independent component analysis of a periodic fMRI complex motor paradigm. Hum. Brain Mapp. 18(2), 111-22.

Moritz, C.H., Carew, J.D., McMillan A.B., and Meyerand M.E., 2005. Independent component analysis applied to self-paced functional MR imaging paradigms. NeuroImage 25(1), 181-192.

Parsons, LM, Fox, PT, Downs, JH, Glass, T, Hirsch, TB, Martin, CC, Jerabek, PA, Lancaster, JL., 1995. Use of implicit motor imagery for visual shape discrimination as revealed by PET. Nature 375(6526), 54-8.

Poline, J., Strother, S.C., Dehaene-Lambertz, G., Egan, G.F., Lancaster, J.L., 2006.Motivation and synthesis of the FIAC experiment: Reproducibility of fMRI results across expert analyses. Hum. Brain Mapp. 27(5), 351-359.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proc Natl Acad Sci U S A. 98(2), 676-82.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu Rev Neuro-

sci. 27, 169-92.

Schmithorst, V.J., and Brown, R.D., 2004. Empirical validation of the triple-code model of numerical processing for complex math operations using functional MRI and group Independent Component Analysis of the mental addition and subtraction of fractions. NeuroImage 22(3), 1414-1420.

Suykens, J.A.K., Van Gestel, T., De Barbanter, J., De Moor, B., and Vanderwalle, J., 2002. Least Squares Support Vector Machines. World Scientific Publishing.

Suhir, E., 1997. Applied Probability for Engineers and scientists. McGraw-Hill.

Suzuki, K, Kiryu, T, Nakada, T, 2001. Fast and precise independent component analysis for high field fMRI time series tailored using prior information on spatiotemporal structure. Hum. Brain Mapp. 15, 54-66.

Svensen, M., Kruggel, F., Benali, H., 2002. ICA of fMRI group study data. NeuroImage 16(3), 551-63.

Thomas, C.G., Richard, A., Harshman, and Menon, R.S., 2002. Noise Reduction in BOLD-Based fMRI Using Component Analysis. NeuroImage 17(3), 1521-1537.

Tucker, DM, Roth, DL, Bair, TB., 1986. Functional connections among cortical regions: topography of EEG coherence. Electroencephalogr Clin Neurophysiol 63, 242–250.

Valente, G., De Martino, F., Balsi, M., Formisano, E., 2005. Optimising ICA using generic knowledge of the sources. IEEE Prime 2005 conference, Lausanne (Switzerland).

van de Ven, V.G., Formisano, E., Prvulovic, D., Roeder, C.H., Linden, D.E., 2004. Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. Hum. Brain. Mapp. 22(3), 165-78.

van de Ven, V.G., Formisano, E., Röder, C.H., Prvulovic, D., Bitter, R.A., Dietz, M.G., Hubl, D., Dierks, T., Federspiel, A., Esposito, F. *et. al.* 2005. The spatiotemporal pattern of auditory cortical responses during verbal hallucinations. NeuroImage 27(3), 644-655.

Vapnik, V., 1979. Estimation of Dependences Based on Empirical Data [in Russian]. Nauka, Moscow. (English translation: 1982, Springer Verlag, New York).

fMRI Independent Components *fingerprints* during Focal Epilepsy 2

Abstract

The General Linear Model (GLM), has been used to analyse simultaneous EEG-fMRI to reveal BOLD changes linked to interictal epileptic discharges (IED) identified on scalp EEG. This approach is ineffective when IED are not evident in the EEG. Data-driven fMRI analysis techniques that do not require an EEG derived model may offer a solution in these circumstances. We compared the findings of independent components analysis (ICA) and EEG-based GLM analyses of fMRI data from eight patients with focal epilepsy. Spatial ICA was used to extract independent components (IC) which were automatically classified as either BOLD-related, motion artefacts, EPI susceptibility artefacts, large blood vessels, noise at high spatial or temporal frequency. The classifier reduced the number of candidate IC by 78%, with an average of 16 BOLD-related IC. Concordance between the ICA and GLM-derived results was assessed based on spatio-temporal criteria. In each patient, one of the IC satisfied the criteria to correspond to IED-based GLM result. The remaining IC were consistent with BOLD patterns of spontaneous brain activity and may include epileptic activity that was not evident on the scalp EEG. In conclusion, ICA of fMRI is capable of revealing areas of epileptic activity in patients with focal epilepsy and may be useful for the analysis of EEG-fMRI data in which abnormalities are not apparent on scalp EEG.

Based on:
Rodionov, R., De Martino, F., Laufs, H., Carmichael, D.W., Formisano, E., Walker, M., Duncan, J.S., Lemieux, L. (2007). Independent Component Analysis of Interictal fMRI in Focal Epilepsy: Comparison with General Linear Model-based EEGcorrelated fMRI. Neuroimage; 38:488-500.

Introduction

In some patients with drug resistant focal epilepsy, surgical resection offers the possibility of seizure control. If the focus is not adequately localized non-invasively, with MRI and scalp EEG, intra-cranial electroencephalography (icEEG) using subdural or intracerebral electrodes may be necessary to define the zone of seizure onset. This procedure is invasive and may fail to identify the epileptogenic zone, in part because of the limited spatial sampling of intracranial electrodes.

Functional MRI enables the non-invasive observation of brain activity with relatively high spatial resolution over the whole brain. EEG-correlated fMRI (EEG–fMRI) has shown promise in epilepsy. Using EEG–fMRI, regions of brain activation and deactivation have been demonstrated in relation to interictal and ictal epileptic discharges, providing a new form of localizing information (Hamandi *et al.*, 2004; Gotman *et al.*, 2006; Laufs and Duncan, 2007). It has been suggested that this type of information could be useful to plan or even to remove the need for intracranial EEG in some cases (Allen *et al.*, 2000; Lemieux *et al.*, 2001). The standard analysis of EEG–fMRI data is based on the identification of interictal epileptiform discharges (IED) on scalp EEG which are used to build a general linear model (GLM) of the fMRI signal changes. In brief, a model is obtained by convolution of the EEG events, which are represented as stick functions or blocks, with a hemodynamic response function; maps showing regions of significant IED-related change are obtained through voxel-wise fitting of the model and application of appropriate statistical thresholds (Lemieux *et al.*, 2001; Benar *et al.*, 2002; Salek-Haddadi *et al.*, 2003; Hamandi *et al.*, 2004; Gotman *et al.*, 2006).

Reliance on scalp EEG for the modelling of BOLD changes throughout the brain is useful to demonstrate hemodynamic changes related to specific EEG patterns but has limitations, because:

(1) The number of EEG events recorded during a 10–40 minute fMRI acquisition must be sufficient for efficient model estimation. The EEG/GLM approach cannot be used when few interictal epileptic discharges (IED) are detected during the fMRI experiment, which is not uncommon (Krakow *et al.*, 1999; Gotman *et al.*, 2004; Salek-Haddadi *et al.*, 2006; Di Bonaventura *et al.*, 2006).

(2) The scalp EEG has limited sensitivity, particularly for activity originating deep in the brain (Tao *et al.*, 2005; Ray *et al.*, 2007), with a resultant bias towards the detection of superficial IED. Furthermore it is well known that epileptic activity originating from deep brain structures in many cases

cannot be recorded on the scalp EEG (Niedermeyer and Lopes da Silva, 2004).

(3) Any epilepsy-related activity which is not detected by scalp EEG will not be modelled, with a potentially significant impact on the technique's sensitivity.

In such circumstances, data-driven fMRI data analysis techniques may provide a way forward (Morgan *et al.*, 2004) as they are not constrained by a fixed hypothesis. This may be particularly beneficial in cases where no hypothesis is available for example in the absence of discharges on the scalp EEG.

ICA is increasingly recognized as a useful fMRI data-driven analysis tool (McKeown *et al.*, 1998; Formisano *et al.*, 2004; Beckmann *et al.*, 2005). Not being reliant on prior hypotheses, ICA of fMRI has the potential to identify a greater proportion of the BOLD signal variations. The main advantage of this method is that it represents the original functional time series as a set of independent components (IC), which may separate meaningful neurophysiological sources and artefacts. However, the lack of a prior hypothesis and the potentially large number of IC generated render interpretation of the results difficult (Formisano *et al.*, 2002; Beckmann and Smith, 2004). Some authors combine ICA with a GLM for fMRI analysis, using IC as GLM regressors (McKeown, 2000; Hu *et al.*, 2005; Mirsattari *et al.*, 2006). Since correlation with an experimental paradigm is a criterion for selection of IC to build a GLM, the core problem of IC interpretation is not tackled in these techniques. The idea of sequential application of spatial and temporal ICA in order to reveal epilepsy-related IC has been suggested, but not evaluated (Chen *et al.*, 2006).

An automated characterization technique has been introduced and implemented to reduce the number of meaningful IC that require interpretation (De Martino *et al.*, 2007). In this method, the classification of patterns as BOLD-like relies on a set of spatial and temporal characteristics derived from data acquired in normal healthy subjects. In the context of epilepsy, it has been suggested that the time course of IED-related BOLD changes may deviate from the canonical shape (Diehl *et al.*, 2003; Salek-Haddadi *et al.*, 2003). However, the IED-related response has been shown to be primarily canonical (Salek-Haddadi *et al.*, 2006; Lemieux *et al.*, 2007) and any deviation is likely to reflect scalp EEG bias rather than coupling or vascular abnormalities (Lemieux *et al.*, 2007; Hawco *et al.*, 2007).

Our goal is to assess the potential of ICA to identify epileptic activity in fMRI datasets acquired in patients with focal epilepsy by comparing the results with the patterns identified using a GLM-based approach (based on EEG abnormalities seen on scalp EEG) in cases with clear electro-clinical localization. We applied

the method of spatial ICA to fMRI data and used an automated classification approach (De Martino *et al.*, 2007), independent of the EEG, to reduce the number of components that are likely to represent epileptic activity. We have chosen to use the classifier as currently trained using data from healthy subjects for this initial study, based on the general normality of the spike-related BOLD response as described above. Then we determined whether the selected components included IED-related spatio-temporal patterns by comparing them with the results of the EEG-based GLM analysis of the same data, resulting in a set of IED-related independent components.

Data

Sixty-three patients with focal epilepsy underwent EEG correlated fMRI (Salek-Haddadi *et al.*, 2006).

Table 1
Patient data and EEG–fMRI GLM results

Case #	Case # in Salek-Haddadi et. al., 2006	Case # in Hamandi et. al., 2005	Aetiology	EEG-fMRI GLM activation localization	Surgery outcome
1	13	B	TLE, L-HS	LH, cuneus	4
2	2	D	MCD	Diffused R mid-temporal and frontal	n/a
3	35	H	MCD	L temporal	n/a
4	12	F	TLE, L-HS	L anterior temporal	2
5	31	C	MCD	R parietal and temporo-parietal	n/a
6	6	A	Chronic encephalitis	L temporal parietal	9
7	23	G	Post-traumatic	R frontal and R parietal	n/a
8	3	E	MCD	L temporal	n/a

The EEG–fMRI GLM activation localization was taken from Salek-Haddadi et al. (2006). Surgery outcome: number of post-surgical years of seizure freedom. Abbreviations: TLE—temporal lobe epilepsy; L-HS—left hippocampus sclerosis; MCD—malformation of cortical development; LH—left hippocampus; R/L—right/left; n/a—no information available.

Eight patients (Table 1) were selected based on the following criteria: (1) clear-cut localization and lateralization on the basis of concordance between clinical seizures, EEG and structural MRI; (2) the localization of the GLM-de-

rived fMRI signal change was concordant with other electro-clinical data (Table 1) (Salek-Haddadi *et al.*, 2006). All patients gave written informed consent (Joint Ethics Committee of the National Hospital for Neurology and Neurosurgery and Institute of Neurology).

BOLD sensitive EPI images (TE/TR 40/3000, 21×5 contiguous, interleaved slices, FOV 24×24 cm, 64×64 matrix) were acquired, with a 1.5 T Horizon Echospeed (General Electric, Milwaukee, USA) MRI scanner with continuous, simultaneous EEG and ECG recording. In seven subjects, 700 scans were acquired. In one case 450 scans were acquired due to a seizure. High-resolution T1 weighted MR images were also acquired (Fast Inversion Recovery [prepared] Spoiled Gradient Recalled [IRSPGR]: TI/TR/TE, 450/15/4.2 (ms), flip angle 25°; 124 1.5 mmthick coronal slices; matrix, 256×192 voxels, 24×18 cm field of view; scan time, 7 min). Patients were asked to rest with their eyes shut and keep their head still.

The EEG was recorded using in-house equipment. Ten channels of common reference EEG in a bi-temporal chain Fp2/Fp1, F8/F7, T4/T3, T6/T5, O2/O1, Fz (ground) and Pz as the reference, according to the international 10:20 system and two channels of ECG were recorded inside the MRI scanner with on line artefact removal (Allen *et al.*, 1998; Allen *et al.*, 2000; Lemieux *et al.*, 2001). The EEG was reviewed off-line and the onset of IED identified by two trained observers who reached consensus on the set of marked IED.

Methods

The fMRI data were analysed using two methods and the results compared: (1) ICA without reference to the simultaneously recorded EEG, and (2) the EEG-based GLM approach (IED correlated fMRI).

ICA of fMRI data

Spatial independent component analysis was performed with Brain Voyager QX software (Brain Innovation, Maastricht, Netherlands). The mathematical details of the ICA for fMRI are described elsewhere (McKeown *et al.*, 1998, 2003). Details of implementation of ICA in BrainVoyager QX are described in Formisano *et al.* (2004).

In summary, the ICA decomposition can be expressed as:

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \qquad (1)$$

where **X** is the measured fMRI signal, **S** the spatial maps of the decomposition and **A**, the time courses defining the relative weighting of the spatial maps

throughout the experiment. We estimated **A** and **S**, using the hierarchical (deflation) mode of the FastICA algorithm (Hyvärinen, 1999), after reduction of the temporal dimension of the data set with principal component analysis to 80 dimensions. This number of dimensions was selected after a preliminary analysis in which we performed consecutive decompositions and verified that the number of IC classified as BOLD was stable and spatiotemporal characteristics of those IC did not change significantly while increasing the number of IC in the decomposition.

Classification of independent components

After decomposition automatic IC classification was applied (see Chapter 1, De Martino *et al.*, 2007) resulting in the following set of labels: (1) the 'BOLD' class, which included components that are thought to consistently reflect task-related, transiently task-related and brain state-related (e.g. default state) neuronal activity; (2) residual motion artefacts; (3) EPI-susceptibility artefacts; (4) physiological noise; (5) noise at high spatial frequency; and (6) noise at temporal high frequency.

The classifier is the result of training with a dataset from healthy volunteers (Chapter 1, De Martino *et al.*, 2007) and is designed to be inclusive rather than restrictive with respect to BOLD components in order to reduce the probability of misclassification of BOLD related IC. Furthermore, it has the ability to reveal stereotypical components of normal brain activity such as the so-called 'default mode' network (Schmithorst and Brown, 2004; Greicius and Menon, 2004) and sensory components reflecting connectivity during rest (Van de Ven *et al.*, 2004). The BOLD components were further classified into the following sub-types by visual inspection:

1. Stereotypical of normal brain activity;
2. Misclassified (False Positives);
3. Other.

The misclassified type contains those with patterns corresponding to one of the following effects (Chapter 1, De Martino *et al.*, 2007): motion, blood vessels, spatially distributed and high frequency noise. We hypothesized that IC classified as 'Other' would contain IC related to IED and form the set of candidate components.

GLM-based analysis of EEG–fMRI data

A GLM-based analysis (Frackowiak *et al.*, 2003) of the fMRI datasets was performed with Brain Voyager QX software based on the identification of IED on the scalp EEG. Two trained observers coded the EEGs together resulting on

a consensus for the entire recordings. Data pre-processing consisted of scan realignment and smoothing using an 8 *mm* kernel (Frackowiak *et al.*, 2003). The GLM comprised HRF-convolved IED-related effects (Salek- Haddadi *et al.*, 2006), motion effects (a combination of Volterra expansion of the 6 scan realignment parameters) (Friston *et al.*, 1996; Lemieux *et al.*, 2007) and scan nulling of head jerks (Salek-Haddadi *et al.*, 2006; Lemieux *et al.*, 2007). Statistical parametric maps (SPM) of activation were obtained by testing for positive BOLD changes using *p=0.05* (Bonferroni corrected). The set of EPI images was co-registered with the T1 volume dataset scaled to Talairach space by means of rigid body transformation for visualization. In these cases, positive BOLD responses were found consistently corresponding to the electro-clinical findings in line with our findings in a larger series (Salek-Haddadi *et al.*, 2006). This is in contrast to the negative BOLD changes, which are generally more remote from the presumed generator. Therefore, only positive IED-related BOLD changes were considered

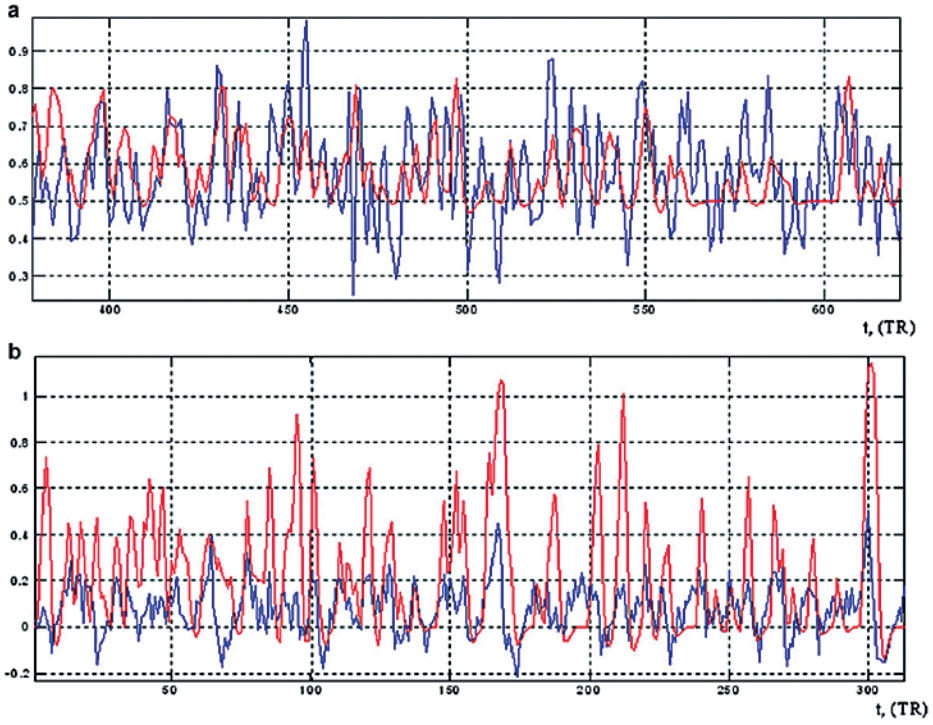

Figure 1: Fragment of the time course of IC (blue) significantly correlated with corresponding IED-related (GLM) regressor (red) for case #4 (a) (R=0.26, p=6.6e−12) and case #7 (b) (R=0.3, p=2.6e−15).

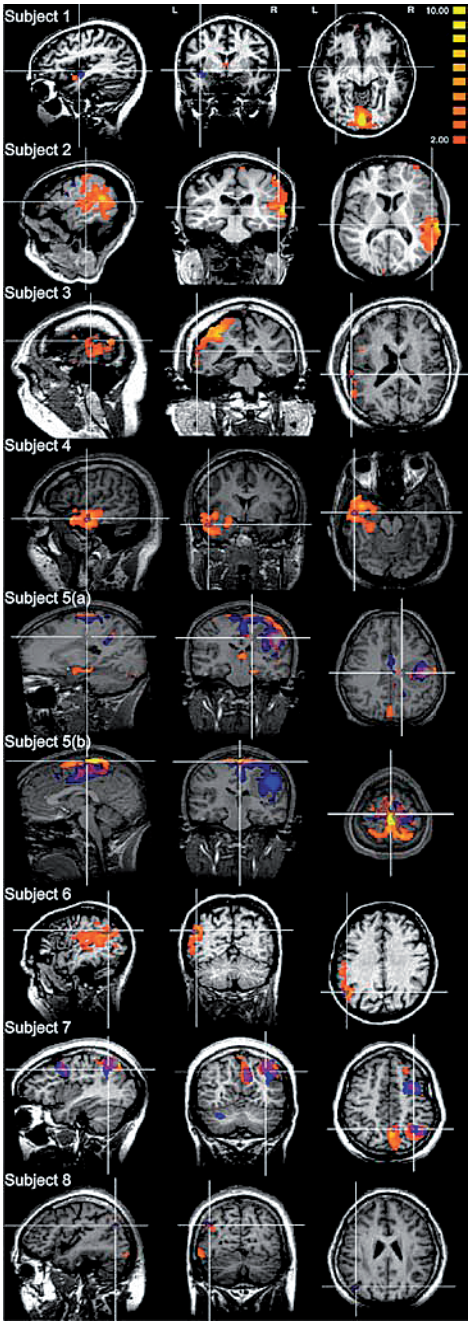

Figure 2: Illustration of SPM (dark blue) and matching IC maps (orange) for all cases, overlap is shown in purple shades. Cursor is located so as to define sections that best illustrate overlap. The thresholds are z=2.0 for the IC maps and p=0.05 (corrected) for the SPM.

in this study.

Comparison of ICA and GLM-based results

The values of the particular IC map represent the relative amount a given voxel is modulated by the activation of that component. To identify significantly contributing voxels, ICA maps were scaled to the spatial z-scores, computed as the number of standard deviations from the map mean (McKeown *et al.*, 1998).

A threshold of *z=2.0* was used for visualization of the IC maps and volumetric comparison of the IC maps and SPM.

Each BOLD component was compared with the GLM solution (positive BOLD changes only). Components that matched the following criteria were classified as matching (IED-related): (1) temporal correlation between component time course and IED-derived regressor with *p=0.01* (Figs. 1 and 2) spatial overlap of the thresholded IC map (*z=2.0*) exceeding 10% of spatial extent of GLM-derived clusters (Fig. 2). The value of 10% of spatial overlap was chosen based on preliminary tests on 3 datasets.

Results

The ICA decompositions are summarized in Table 2. The number of BOLD components ranged between 11 and 24 (mean: 16, median: 15) representing an average of 22% of the total (Table 2). In six cases one, and in one case, two of the candidate components matched the GLM-derived result according to the set criteria. In one case, #4, no BOLD component was found to meet the matching criteria (Table 2, Fig. 2). In accordance with previous studies of the resting state in healthy subjects, components characteristic of the so-called resting state networks (RSN) were found consistently in the group studied. These involved sensory–motor, visual and auditory areas and the so-called default mode network (DMN: precuneus and frontooccipital areas), the latter illustrated in Fig. 3.

Table 2
ICA-GLM comparison

Case #	# BOLD IC s	# GLM-concordant BOLD IC	IC-SPM overlap % (# of voxels)	Significance of R
1	11	1	41.5 (35)	2.39e-6
2	9	1	18.5 (67)	0.009
3	24	1	60 (9)	0.01
4	21	0	-	-
5	20	2	41 (500); 23 (280)	3.4e-9 1.7e-7
6	13	1	78 (207)	2.0e-20
7	17	1	25 (671)	2.6e-15
8	11	1	15 (2)	2.6e-4

The volume of overlap between IC maps and SPMs is expressed both as a percentage of the SPM activation cluster. The significance of the correlation between IC time course and spike regressor is expressed as the probability of getting the same value of correlation coefficient by chance.

Case reports

In accordance with the selection criteria used for this study, the result of the GLM analysis was concordant with the presumed or confirmed focus for all cases (Salek-Haddadi *et al.*, 2006).

Case 1

Patient with left hippocampal sclerosis and left temporal IEDs. The GLM analysis revealed a left temporal activation with an additional smaller activation cluster in the occipital region. The single matching IC consisted of a cluster in the

left hippocampus and a larger one in the occipital region, either overlapping or adjacent to the GLM pattern (Fig. 2, subject 1).

Case 2

Patient with an extensive malformation of cortical development in the right hemisphere, involving predominantly the parietal lobe but extending to the occipital and frontal lobes, plus focal atrophy of the left parietal lobe. GLM analysis revealed an area of right frontal and mid-temporal activation linked to right temporal IEDs. The GLM analysis for this patient has a lower level of significance possibly due to the large degree of motion (Salek-Haddadi *et al.*, 2006). One IC matched the GLM result, covering the parietal part of the lesion and corresponding clusters of GLM activation (Fig. 2, subject 2).

Case 3

Patient with left parietal polymicrogyria, left hemisphere atrophy and left hippocampal sclerosis. During EEG–fMRI there were frequent left anterior temporal spikes which were linked to a small cluster of left temporal activation (Fig. 2, subject 3). The matching IC corresponded to the GLM activation plus most of the cortical lesion.

Case 4

Patient with left hippocampal sclerosis and left anterior temporal spikes. EEG–fMRI revealed lateral temporal lobe activation. No IC was found that satisfied the matching criteria.

Case 5

Patient with two large heterotopic nodules, frontoparietocentral and medial parietal. During EEG–fMRI there were frequent right central spikes or varying amplitudes plus slow waves. GLM activation clusters were observed within both nodules. Two IC matching the concordance criteria were found: one in the nodules (Fig. 2, subject 5 (a)) and the other under the vertex (Fig. 2, subject 5 (b)).

Case 6

Patient had left hemisphere chronic encephalitis of adult onset. The GLM activation was linked to high-amplitude left temporal sharp-wave discharges. A single matching IC was found covering most of the GLM-derived activation (Fig. 2, subject 6).

Case 7

Patient had a post-traumatic right middle frontal gyrus scar. EEG–fMRI revealed three distinct clusters of activation linked to runs of polyspike-wave activity. A single matching IC was found with excellent correspondence to the GLM result (Fig. 2, subject 7).

Case 8

Patient had widespread predominantly posterior band and nodular heterotopia and bursts of posterior temporal/occipital discharges with left-sided emphasis. EEG–fMRI revealed a small cluster of left temporal-occipital activation concordant with structural MRI and interictal EEG. Matching IC consisted of a small cluster overlaying the GLM result (Fig. 2, subject 8).

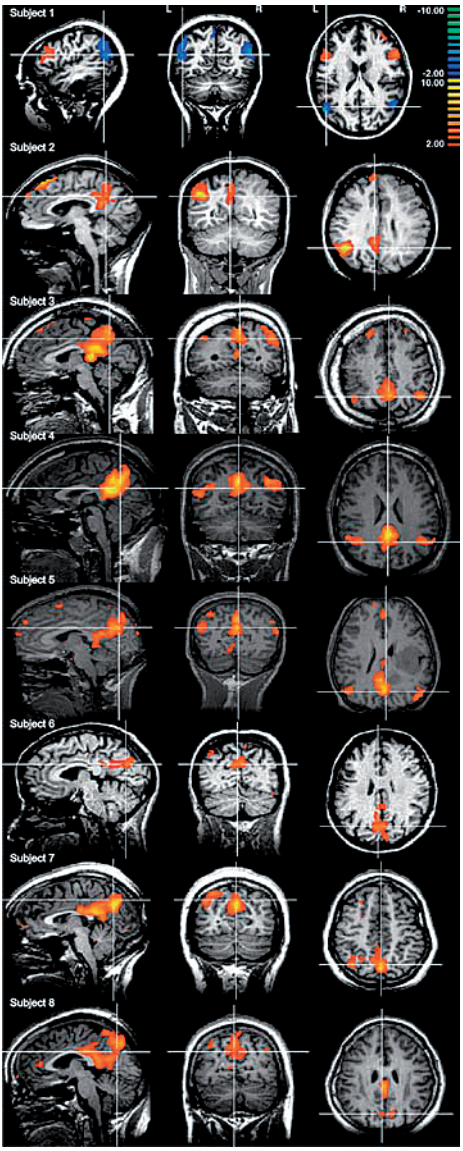

Figure 3: Illustrative IC maps of components corresponding to resting state networks (Greicius *et al.*, 2003; Fransson, 2005). Each IC map is visualized at a threshold of zN2.0; the numbers for each set of projections correspond to # of case ID in Tables 1 and 2.

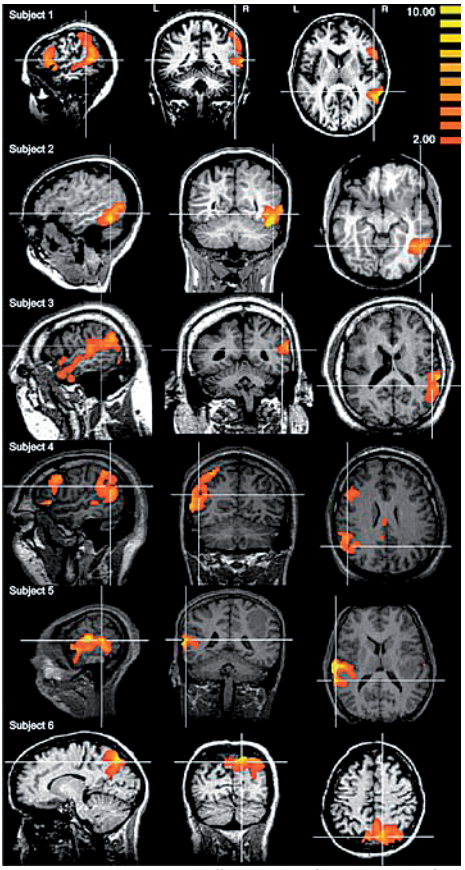

Figure 4: Representative illustration of IC requiring further investigation and interpretation for cases 1–6. Visualization threshold zN2.0. Case 1: contralateral to the GLM result; case 2: posterior to the GLM result; case 3: contralateral to the lesion and GLM result; case 4: posterior to the GLM result; case 5: contralateral to the lesion and GLM result; case 6: contralateral to the lesion and GLM result. The numbers for each set of projections correspond to # of case in Tables 1 and 2.

Discussion

We analysed the results of ICA decomposition of fMRI data in patients with focal epilepsy in whom EEG–fMRI had revealed GLM-derived activation patterns judged to be concordant with the epileptic focus. Using automated component classification, in 7/8 cases it was possible to identify one or two IC which corresponded to GLM activations.

Methodological issues

The classifier we used was trained on data from healthy subjects in a block-designed visual experiment, which may be sub optimal for application in patient data where spontaneous brain activity is recorded without external stimulation. Visual inspection of the components classified as BOLD by expert observers revealed that 50% of these did not correspond to BOLD effects but rather to motion, big blood vessels and high frequency noise. This represents a trade-off in favour of classifier sensitivity over specificity.

The classifier failed to detect a matching BOLD-related component in a single case (#4). In this case, inspection of all IC by expert observers revealed a component that qualitatively matched the GLM result that had been classified as temporal high frequency noise. This case was characterized by the highest degree of correlation between motion expressed as displacement in 3D space calculated with realignment parameters and the GLM regressor ($R=-0.3$, $p=0.01$). It should be noted that typical patterns for motion and temporal high frequency noise have similar values for the high frequency parameter of the fingerprint (De Martino *et al.*, 2007) which caused misclassification of the IC. Fragments of the time course of the selected IC and IED-based regressor are shown in Fig. 1a ($R=0.25$, $p=6.6e-12$).

In patients, fMRI data are often contaminated by motion and so caution is required when analysing and interpreting results, especially when applying data-driven approaches. For example, in case # 2 with 108/450 scans modelled as head jerks, there was considerable overlap between a number of motion-related IC and the GLM result. The presence of motion-related noise risks masking genuine signal changes, particularly when temporally correlated with the events of interest. In case #4, we concluded that the presence of motion artefact (highly significant correlation of IED-based regressor and integral motion parameter $R=-0.3$, $p=2.6e-15$) caused misclassification of the GLM-concordant IC (Fig. 2, subject 4) due to high temporal frequency contamination.

We evaluated each BOLD IC based on spatial overlap with GLM activation pattern and temporal correlation with the GLM IED-derived regressor (Fig. 1b).

Although the choice of spatial overlap and temporal correlation threshold values can appear arbitrary the results provide some reassurance that this approach is useful.

Interpretation of BOLD components

One of the key elements of our approach was a data reduction strategy to reduce the number of IC to be matched with the results. The first step of this was the automatic classification of IC as BOLD components. The second step was the visual identification of IC with spatial patterns typical of Resting State Networks: visual, auditory, sensory–motor and so-called Default Mode, in line with previous studies of resting state fMRI data using data-driven techniques (Seifritz *et al.*, 2002; Beckmann *et al.*, 2005; Fransson, 2005; Fox *et al.*, 2005; Aragri *et al.*, 2006; Laufs *et al.*, 2006). We note the asymmetry in the DMN in case #2 which may reflect pathology of the right hemisphere (Fig. 3, subject 2).

The remaining BOLD IC formed the set of components which comprised epilepsy-related components (in 7/8 cases) and components for which we could not find an interpretation. Examples of such IC are shown in Fig. 4. Some of the uninterpreted IC (i.e. that did not satisfy both the spatial and temporal matching criteria) had a time course that was significantly correlated with the IED-derived GLM regressor, providing evidence that those IC could describe some parts of the spatial distribution of epileptogenic network. This is equivalent to sub-threshold activity from the point of view of the GLM and may therefore simply reflect differences in the way significance is established for the two methods. As an additional analysis, we studied the temporal relationship between the 'Other' (potentially epileptic) IC and the relevant GLM regressor as a function of relative time lag. This has revealed significant correlation in 5 of the 7 cases in which potentially epileptic components were identified, with peak correlation at time lags ranging between −1 TR and 2 TR. For example, in case #5 the correlation peaked at a time lag of +1 TR, suggesting delayed BOLD changes contralateral to the presumed focus in relation to the ipsilateral BOLD effect (as revealed by the GLM) (Fig. 5). The significance of such lags, which may in part

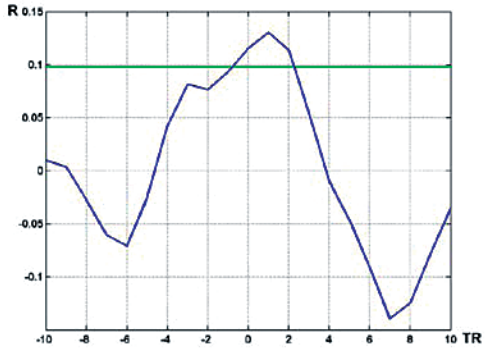

Figure 5: Illustration of correlation function between IED-derived regressor and time course of an IC interpreted as 'Other' (case #5). The map of the IC is shown in Fig. 4, subject 5. The level of significance of correlation is R=0.98 (green line). Here peak correlation at +1 TR means a delay of the IC time course in relation to IED-derived regressor.

reflect normal variability in the hemodynamic response which may be captured using a more flexible HRF basis set, remains to be studied. Nonetheless, the additional components identified in this study illustrate an interesting aspect of data-driven analyses: that of possible added value compared to the GLM-based analysis by removing the need for tight temporal coupling between scalp EEG event and hemodynamic fluctuations. Validation of such components would require correlation with the complete characterization of the epileptogenic network as may be obtained with intracranial EEG. Our assumption is that epileptiform activity recorded on the scalp, and therefore which originates in the superficial cortex, does not fundamentally differ from activity that is generated deeper inside the brain (and that is not reflected on scalp EEG) at least hemodynamically. No additional potentially epilepsy-related IC were found for cases 7 and 8.

As mentioned previously, in case #4 no candidate IC matched the EEG-derived pattern. Inspection of the components classified as non-BOLD revealed one for which both the spatial and temporal correspondence criteria were satisfied (Fig. 2.4): spatial overlap between the IC map and the SPM was 93% (96 voxels) of the SPM volume, the correlation of the IC time course and the IED-based regressor was R=0.26 with a very high significance ($p=6.6e{-}12$, Fig. 1a). As suggested above, this IC was probably contaminated by motion.

One of the two concordant IC in case #5 (Fig. 2.5b) was observed in all datasets; in principle, there are two possible explanations namely that the component is commonly linked to resting state activity or to epilepsy. Its location close to the sagittal sinus, and distribution pattern suggest the former.

Biological and clinical significance

The nature of the selected cases allows us to conclude that the matching of ICA and GLM results confirms a general concordance between the identified component(s) and the generator of the IED. Patients 1, 4 and 6 were seizure free 4, 2 and 9 years respectively after left temporal lobe resection. Comparison of the fMRI and post-operative structural MRI revealed that the activated areas were resected, providing strong evidence that ICA was capable of highlighting areas of brain involved in epileptic activity in those cases.

It is important to note that none of the GLM and IC results matched perfectly. For example, only a small degree of spatial overlap between GLM results and IC maps does not preclude functional correspondence. The possible explanations for differences in spatial distribution and time course include: differences in the underlying mathematical assumptions between the models; differences in criteria for selecting significant voxels (thresholding); bias of the GLM analysis due to limited EEG sensitivity, EEG bias towards more superficial activity and subjectiv-

ity of the EEG interpretation. Thus in case #4 when increasing the threshold up to z=3.0 and further to z=4.0 there was still significant spatial overlap with the SPM, at 85% and 52% of SPM volume correspondingly. The spatial configuration of the IC maps and the SPM appeared more similar to each other with these higher thresholds than for the case of z=2.0 shown in Fig. 2.4.

Future work

The ability of ICA to blindly decompose the time series in a set of independent (and physiologically meaningful) components makes it an attractive tool to formulate spatial hypotheses with respect to the source of IED. ICA alone, however, does not allow distinguishing IED-related activity from other BOLD related activation patterns (e.g. default mode and resting state patterns).

The next step would be to perform ICA of fMRI in patients in whom no EEG-derived model is available due to the absence of epileptic activity in the EEG data acquired during EEG–fMRI acquisition. In such cases the GLM approach is not applicable and it will be necessary to develop new methods based on non temporal priors in order to determine which IC are epilepsy-related. Retraining or refinement of the classifier according to resting state data in healthy subjects or epileptic patients may allow increased specificity. Additional information such as longer duration EEG recordings outside the scanner, intra-cranial EEG and surgical data may also be used. This could lead to a methodology for further reducing the number of candidate epileptiform components. Validation could be performed based on comparison with information derived from intra-cranial EEG data and post-surgical outcome.

Conclusions

We found that ICA of interictal fMRI is capable of revealing areas of epileptic activity in patients with well-characterized focal epilepsy. The automatic IC classification method used in this study resulted in a limited number of IC, hypothesized to correspond to epileptic activity. In all cases studied at least one of the IC was consistent with EEG-based BOLD activations and due to the nature of the patients selected represent epileptic regions that generate IED. In a number of cases, IC were revealed that did not correspond to the EEG-derived result but that could be the basis for further investigation, and in particular by comparison with results of other diagnostic modalities, to assess the method's ability to identify pathological changes not linked to scalp EEG abnormalities.

Acknowledgments

This work was funded by the Medical Research Council (grant number G0301067; RR and DC), the Wellcome Trust (grant number 067176; LL), the Bundesministerium für Bildung und Forschung (HL) and Deutsche Forschungsgemeinschaft grant LA 1452/3-1 (HL). Thanks to Mark Symms for MRI developments, Philippa Bartlett and Jane Burdett for MRI scanning. The support of the National Society for Epilepsy (UK) and Brain Innovation (Netherlands) is acknowledged.

References

Allen, P.J., Polizzi, G., Krakow, K., Fish, D.R., Lemieux, L., 1998. Identification of EEG Events in the MR Scanner: the Problem of Pulse Artefact and a Method for Its Subtraction. Neuroimage 8, 229–239.

Allen, P.J., Josephs, O., Turner, R., 2000. A method for removing imaging artifact from continuous EEG recorded during functional MRI. Neuro- Image 12, 230–239.

Aragri, A., Scarabino, T., Seifritz, E., Comani, S., Cirillo, S., Tedeschi, G., Esposito, F., Di Salle, F., 2006. How does spatial extent of fMRI datasets affect independent component analysis decomposition? Hum. Brain Mapp. 27, 736–746.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imag. 23 (2), 137–152.

Beckmann, C.F., DeLuca, M., Devlin, J.T., Smith, S.M., 2005. Investigations into resting-state connectivity using independent component analysis. Philos. Trans. R. Soc. Lond., B Biol. Sci. 360 (1457), 1001–1013.

Benar, C.G., Gross, D.W., Wang, Y.H., Petre, V., Pike, B., Dubeau, F., Gotman, J., 2002. The BOLD response to interictal epileptiform discharges. NeuroImage 17, 1182–1192.

Chen, H., Yao, D., Lu, G., Zhang, Z., Hu, Q., 2006. Localization of latent epileptic activities using spatio-temporal independent component analysis of fMRI data. Brain Topogr. 19 (1–2), 21–28.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., Formisano, E., 2007. Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. Neuro- Image 34 (1), 177–194.

Di Bonaventura, C., Vaudano, A., Carnì, M., Pantano, P., Nucciarelli, V., Garreffa, G., Maraviglia, B., Prencipe, M., Bozzao, L., Manfredi, M., Giallonardo, A., 2006. EEG/fMRI study of ictal and interictal epileptic activity: methodological issues and future perspectives in clinical practice. Epilepsia 47 (s5), 52–58.

Diehl, B., Salek-Haddadi, A., Fish, D.R., Lemieux, L., 2003. Mapping of spikes, slow waves, and motor tasks in a patient with malformation of cortical development using simultaneous EEG and fMRI. Magn. Reson. Imaging

21 (10), 1167–1173.

Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Price, C.J., Zeki, S., Ashburner, J., Penny, W.D., 2003. Human Brain Function, 2nd edition. Academic Press.

Fransson, P., 2005. Spontaneous low-frequency BOLD signal fluctuations: an fMRI investigation of the resting-state default mode of brain function hypothesis. Hum. Brain Mapp. 26 (1), 15–29.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., Turner, R., 1996. Movement-related effects in fMRI time-series. Magn. Reson. Med. 35 (3), 346–355.

Formisano, E., Esposito, F., Kriegeskorte, N., Tedeschi, G., Di Salle, F., Goebel, R., 2002. Spatial independent component analysis of functional magnetic resonance imaging time-series: characterization of the cortica components. Neurocomputing 49, 241–254.

Formisano, E., Esposito, F., Di Salle, F., Goebel, R., 2004. Cortex-based independent component analysis of fMRI time-series. Magn. Reson. Imaging 22 (10), 1493–1504.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U. S. A. 102 (27), 9673–9678.

Greicius, M.D., Menon, V., 2004. Default-mode activity during a passive sensory task: uncoupled from deactivation but impacting activation. J, Cogn. Neurosci. 16 (9), 1484–1492 (Nov).

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc. Natl. Acad. Sci. U. S. A. 100 (1), 253–258.

Gotman, J., Benar, C.G., Dubeau, F., 2004. Combining EEG and FMRI in epilepsy: methodological challenges and clinical results. J. Clin. Neurophysiol. 21 (4), 229–240.

Gotman, J., Kobayashi, E., Bagshaw, A.P., Benar, C.G., Dubeau, F., 2006. Combining EEG and fMRI: a multimodal tool for epilepsy research. J. Magn. Reson. Imaging 23 (6), 906–920.

Hamandi, K., Salek-Haddadi, A., Fish, D.R., Lemieux, L., 2004. EEG/functional MRI in epilepsy: the Queen Square experience. J. Clin. Neurophysiol. 21 (4), 241–248.

Hamandi, K., Salek Haddadi, A., Liston, A., Laufs, H., Fish, D.R., Lemieux, L., 2005. fMRI temporal clustering analysis in patients with frequent interictal epileptiform discharges: comparison with EEGdriven analysis. NeuroImage 26 (1), 309–316.

Hawco, C.S., Bagshaw, A.P., Lu, Y., Dubeau, F., Gotman, J., 2007. BOLD changes occur prior to epileptic spikes seen on scalp EEG. NeuroImage 35 (4), 1450–1458.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. 10 (3), 626–634.

Hu, D., Yan, L., Liu, Y., Zhou, Z., Friston, K., Tan, C., Wu, D., 2005. Unified SPM-ICA for fMRI analysis. NeuroImage 25, 746–755.

Krakow, K., Woermann, F.G., Symms, M.R., Allen, P.J., Lemieux, L., Barker, G.J., Duncan, J.S., Fish, D.R., 1999. EEG-triggered functional MRI of interictal epileptiform activity in patients with partial seizures. Brain 122, 1679–1688.

Laufs, H., Duncan, J.S., 2007. Electroencephalography/functional MRI in human epilepsy: what it currently can and cannot do. Curr. Opin. Neurol. 20 (4), 417–423.

Laufs, H., Holt, J.L., Elfont, R., Krams, M., Paul, J.S., Krakow, K., Kleinschmidt, A., 2006. Where the BOLD signal goes when alpha EEG leaves. NeuroImage 31, 1408–1418.

Lemieux, L., Salek-Haddadi, A., Josephs, O., Allen, P., Toms, N., Scott, C., Krakow, K., Turner, R., Fish, D.R., 2001. Event-related fMRI with simultaneous and continuous EEG: description of the method and initial case report. NeuroImage 14, 780–787.

Lemieux, L., Salek-Haddadi, A., Lund, T.E., Laufs, H., Carmichael, D., 2007. Modelling large motion events in fMRI studies of patients with epilepsy. Magn. Reson. Imaging 25 (6), 894–901.

McKeown, M.J., 2000. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. NeuroImage 11, 24–35.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Mapp. 6, 160–188.

McKeown, M.J., Hansen, L.K., Sejnowski, T.J., 2003. Independent component analysis of functional MRI: what is signal and what is noise? Curr. Opin.

Neurobiol. 13, 620–629.

Mirsattari, S.M., Wang, Z., Ives, J.R., Bihari, F., Leung, L.S., Bartha, R., Menon, R.S., 2006. Linear aspects of transformation from interictal epileptic discharges to BOLD fMRI signals in an animal model of occipital epilepsy. NeuroImage 30, 1133–1148.

Morgan, V.L., Price, R.R., Arain, A., Modur, P., Abou-Khalil, B., 2004. Resting functional MRI with temporal clustering analysis for localization of epileptic activity without EEG. NeuroImage 21, 473–481.

Niedermeyer, E., Lopes da Silva, F., 2004. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, 1st edition. Williams & Wilkins.

Ray, A., Tao, J.X., Hawes-Ebersole, S.M., Ebersole, J.S., 2007. Localizing value of scalp EEG spikes: a simultaneous scalp and intracranial study. Clin. Neurophysiol. 118 (1), 69–79.

Salek-Haddadi, A., Friston, K.J., Lemieux, L., Fish, D.R., 2003. Studying spontaneous EEG activity with fMRI. Brain Res. Rev. 43, 110–133.

Salek-Haddadi, A., Diehl, B., Hamandi, K., Merschhemke, M., Liston, A., Friston, K., Duncan, J.S., Fish, D.R., Lemieux, L., 2006. Hemodynamic correlates of epileptiform discharges: an EEG–fMRI study of 63 patients with focal epilepsy. Brain. Res. 1088 (1), 148–166.

Schmithorst, V.J., Brown, R.D., 2004. Empirical validation of the triple-code model of numerical processing for complex math operations using functional MRI and group independent component analysis of the mental addition and subtraction of fractions. NeuroImage 22, 1414–1420.

Seifritz, E., Esposito, F., Hennel, F., Mustovic, H., Neuhoff, J.G., Bilecen, D., Tedeschi, G., Scheffler, K., Di Salle, F., 2002. Spatiotemporal pattern of neuronal processing in the human auditory cortex. Science 297, 1706–1708.

Tao, J.X., Ray, A., Hawes-Ebersole, S., Ebersole, J.S., 2005. Intracranial EEG substrates of scalp EEG interictal spikes. Epilepsia 46 (5), 669–676.

Van de Ven, V.G., Formisano, E., Prvulovic, D., Roeder, C.H., Linden, D.E., 2004. Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. Hum. Brain Mapp. 22 (3), 165–178.

Mapping fMRI patterns using Support Vector Machines and Recursive Feature Elimination

3

Abstract

In functional brain mapping, pattern recognition methods allow detecting multivoxel patterns of brain activation which are informative with respect to a subject's perceptual or cognitive state. The sensitivity of these methods, however, is greatly reduced when the proportion of voxels that convey the discriminative information is small compared to the total number of measured voxels. To reduce this dimensionality problem, previous studies employed univariate voxel selection or region of interest-based strategies as a preceding step to the application of machine learning algorithms.

In this chapter we describe a strategy for classifying functional imaging data based on a multivariate feature selection algorithm, Recursive Feature Elimination (RFE) that uses the training algorithm (support vector machine) recursively to eliminate irrelevant voxels and estimate informative spatial patterns. Generalization performances on test data increases while features/voxels are pruned based on their discrimination ability.

Using simulated fMRI data, we evaluate RFE in terms of sensitivity of discriminative maps (Receiver Operative Characteristic analysis) and generalization performances and compare it to previously used univariate voxel selection strategies based on activation and discrimination measures.

Furthermore, we apply our method to high resolution data from an auditory fMRI experiment in which subjects were stimulated with sounds from four different categories. With these real data, our recursive algorithm proves able to detect and accurately classify multivoxel spatial patterns, highlighting the role of the superior temporal gyrus in encoding the information of sound categories. In line with the simulation results, our method outperforms univariate statistical analysis and statistical learning without feature selection.

Based on:

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E. (2008). Combining multivariate voxel selection and Support Vector Machines for mapping and classification of fMRI spatial patterns. (submitted)

Introduction

Machine Learning and pattern recognition techniques are being increasingly used in fMRI data analysis. These methods allow detecting subtle, non-strictly localized effects that may remain invisible to the conventional analysis with univariate statistics (Norman *et al.*, 2006, Haynes *et al.*, 2006). In contrast to these latter approaches, machine learning techniques take into account the full spatial pattern of brain activity, measured simultaneously at many locations, and exploit the inherent multivariate nature of fMRI data.

The application of machine learning techniques to fMRI has been referred to as multivoxel pattern analysis (MVPA) and it generally entails four steps (Norman *et al.*, 2006). First, the set of voxels that will enter the multivariate analysis is selected. With respect to this, the analysis may be *massively* multivariate and consider all brain voxels simultaneously (whole-brain approach, Mourao-Miranda *et al*., 2005) or may be limited to a subset of voxels from one region-of-interest (ROI) (Cox and Savoy, 2003, Haynes *et al.*, 2005, Kamitani *et al.*, 2005), in which case the dimensionality of the multivariate space is greatly reduced. Second, stimulus-evoked brain activity is represented as a point in a multidimensional space, i.e. as the pattern of intensity values at selected voxels (multivoxel patterns, MVP). In order to represent the brain response to a stimulus or cognitive state any estimate of activation at the selected voxels can be used, such as the intensity at a single acquisition volume (TR) (Haynes *et al.*, 2005, Mourao-Miranda *et al.*, 2005) or the average intensity in multiple TRs (Kamitani *et al.*, 2005, Mourao-Miranda *et al.*, 2006). Third, using a subset of trials, a classifier is trained and the optimal separating boundary (hypersurface) between different conditions in this multidimensional space is defined. Several methods including Support Vector Machines (SVMs) (Cox and Savoy, 2003, Mitchell *et al.*, 2004, Mourao-Miranda *et al.*, 2005, LaConte *et al.*, 2005, Kamitani *et al.*, 2005), linear discriminant analysis (LDA) (Haynes *et al.*, 2005, Kriegeskorte *et al.*, 2006), and Gaussian Naïve Bayes (GNB) (Mitchell *et al.*, 2004) classifiers have been used for this purpose. During training, a map coding for the relative contribution of each voxel to the discrimination of conditions (discriminative map) can be directly obtained for all linear classifiers (Mourao-Miranda *et al.*, 2005). Fourth, the capability of the trained classifier to accurately discriminate the experimental conditions when presented with new data (i.e. trial responses not used during training) is tested (generalization).

This chapter deals with issues concerning the first point, i.e. the initial selection of the set of voxels, with the aim of optimizing the performance of the multivoxel pattern analysis. For consistency with the pattern recognition literature the

voxels of an fMRI data set are also referred to as “features”.

Whole-brain approaches are appealing in that they do not require *a priori* hypothesis on the location of the relevant voxels, which can be determined *post hoc* from the *discriminative maps*. These approaches seem most appropriate when the discrimination of perceptual or cognitive states is reflected by widely distributed activation patterns that extend and include various and separated brain regions. However, whole-brain approaches may be problematic when the aim of the analysis is the fine-grained discrimination between perceptual states (Haynes *et al.*, 2005, Kamitani *et al.*, 2005). In fact, in these cases the proportion of voxels that convey the discriminative information is expected to be small and thus whole brain approaches seem sub-optimal. Machine learning algorithms are known to degrade their performances when faced with many irrelevant features (overfitting, Kohavi *et al.*, 1997, Guyon *et al.*, 2003, Norman *et al*., 2006), especially, when the number of training samples is rather limited as in typical fMRI studies. Thus selection of an adequate subset of features/voxels is of critical importance in order to obtain classifiers with good generalization performance.

Restricting the multivariate analysis to an anatomically or functionally pre-defined subset of voxels can be seen as a solution to this feature selection problem. This solution is affected by all limitations of ROI-based approaches, which only allow testing a limited set of spatial hypotheses and cannot be used when the aim of the study is the localization of those voxels forming discriminative patterns. An interesting alternative is the local multivariate search approach proposed by Kriegeskorte *et al.*, (2006). This method relies on the assumption that the discriminative information is encoded in neighbouring voxels within a “searchlight” of specified radius. Such locally-distributed analysis might be, however, sub optimal when no hypothesis is available on the size of the neighbourhood and might fail to detect discriminative patterns jointly encoded by distant regions (e.g. bilateral activation patterns).

The main limitation of whole-brain MVPA is its computational complexity since the number of voxels is very large in comparison to the number of trials in a typical fMRI acquisition (Norman *et al.*, 2006). In pattern recognition approaches, feature selection strategies are usually employed prior to the analysis in order to reduce the dimensionality and to preserve sensitivity to small effects. In previous neuroimaging applications, machine learning algorithms have been combined with univariate feature selection strategies (Mitchell *et al.*, 2004, Mourao-Miranda *et al.*, 2006). Both the activation level (F-test) or the discrimination ability (t-test) have been used as univariate ranking criteria for voxel selection. Any such method of voxel selection, though, does not consider the inherent multivariate nature of the fMRI data.

Multivariate feature selection strategies can be summarized in three categories, multivariate filters, wrappers and embedded methods (for a review see Kohavi *et al.*, 1997 and Guyon *et al.*, 2003). Filter methods are applied previous to the classifier and thus do not make use of the classifier performance to evaluate the feature subset. Wrappers and embedded methods, on the other end, use the classifier to find the best feature subset. Wrappers consider the classifier as a black box and make use of different search engines in the feature space to find the subset that maximizes generalization performances. Embedded methods instead incorporate feature selection as a part of the training process.

Here we consider an approach to fMRI MVPA that ensures high sensitivity to fine grained spatial discriminative patterns, while preserving the appealing properties of whole-brain analysis. This approach combines a wrapper method (Recursive Feature Elimination) and SVMs to perform fMRI MVPA. Recursive Feature Elimination (RFE) has been compared to other multivariate feature selection strategies (Rakotomamonjy, 2003) and has been successfully applied to gene selection and sample classification in combination with Support Vector Machine classifiers (Guyon *et al.*, 2002). The recursive nature of the algorithm makes RFE computationally feasible in fMRI MVPA where the number of features can be up to 300.000 cortical voxels, as in the case of whole brain high resolution (2 x 2 x 2 mm^3) acquisitions. In a recent publication Hanson *et al.* (in press) introduced the combination of RFE and SVMs to fMRI multivoxel pattern recognition analysis. The authors showed that removing iteratively irrelevant voxels improves generalization performances in discriminating visual stimuli (Faces and Houses) during two different tasks (1-back recognition detection task, oddball).

In this chapter, we evaluate and compare the performances of RFE, of different univariate feature filter methods (activation and discrimination based), and of their combination on simulated fMRI data. For each method, sensitivity analysis (ROC analysis) and generalization performance are computed at different levels of functional signal-to-noise (SNR, activation level with respect to the baseline) and functional contrast-to-noise (CNR, differences between activation levels in two conditions). We show that, especially in the case of low SNR and/or CNR, the combination of univariate activation-based voxel reduction and RFE outperforms all other methods.

We also apply our method to a real data set obtained in an experiment with auditory stimulation in which sounds from four different categories were presented. The results of the analysis on this fMRI data set confirm the expectations from the simulations and show that the combination of activation based univariate feature selection and RFE provides the highest generalization performance.

Methods

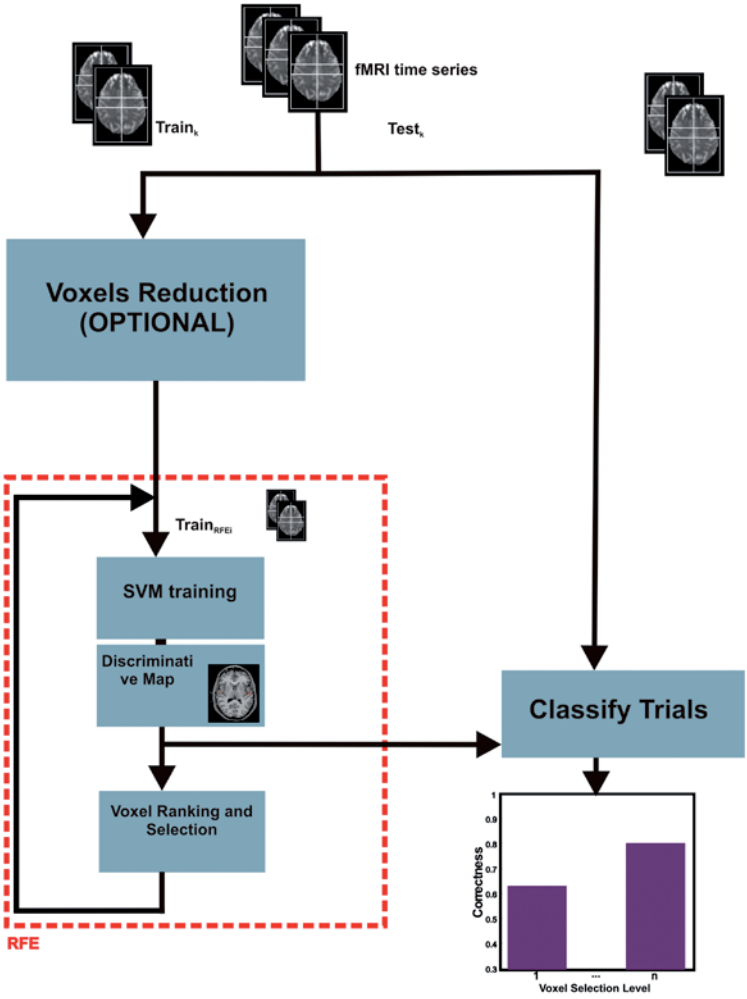

Figure 1: General description of the proposed SVM/RFE iterative procedure to brain mapping. After single trial response estimation functional time series are divided in training and test data sets. An optional step of voxel reduction can be performed prior to RFE using only the training data. For each voxel selection level the recursive procedure (RFE; red dashed box in figure) consists of two steps. First an SVM classifier is trained using the current set of voxels. Second a set of voxels is discarded according to their discriminative weights as estimated during training. Test data are classified at each iteartion and generalization performances are assessed.

General description of the approach

Figure 1 illustrates schematically the proposed approach. Trials from the fMRI time-series are divided into training and test set; the latter data is only used to assess generalization performance. The main processing stage is the multi-variate feature selection using Recursive Feature Elimination (RFE, red dashed box in Figure 1). At each iteration, RFE includes two steps. First, a subset of the

training data ($Train_{RFEi}$) is used to train an SVM classifier. Second, discrimination weights obtained from the SVM training are ranked and voxels corresponding to the smallest ranking are discarded. Voxels with the highest discriminative values are used for training in the next iteration. These two steps are repeated multiple times, thus generating a set of discriminative maps which progressively include a smaller number of voxels. Each of the times, the accuracy of classification corresponding to the current discriminative map is assessed using the external test trials. Assessment of the performance is obtained using an *n*-fold cross-validation scheme, i.e. repeating the whole analysis with different splitting of training and test data sets ($Train_k$, $Test_k$). Estimates of accuracy and discriminative maps for each voxel selection level are obtained by averaging across different folds.

Figure1 also indicates that prior to RFE an additional, preliminary step of univariate voxel reduction can be used. This step consists in selecting a subset of voxel, based on univariate statistics computed on the training data ($Train_k$).

Single trial response estimation.

We estimated the multivoxel pattern of intensities forming the input to the SVM classifier in the following way. At each stimulus presentation, a trial t (t = 1…T), is formed considering N_{pre} and N_{post} temporal samples (before and after stimulus onset respectively) of the pre-processed (see below) time course of activity. A trial estimate of the response at every voxel v (v = 1,...V) is then obtained by fitting a General Linear Model (GLM) with one predictor coding for the trial response and one linear predictor accounting for a within-trial linear trend. The trial-response predictor is obtained by convolution of a boxcar with a double-gamma hemodynamic response function (HRF) (Friston *et al.*, 1998). At every voxel, the corresponding regressor coefficient (beta) is taken to represent the trial response.

To account for BOLD response variability we repeated the estimation procedure (at each voxel) changing the time-to-peak (four to six seconds) of the modelled HRF response. The best-fitting beta (minimum *p-value*) was selected as representative for the trial response.

The outlined procedure is designed for slow event related designs or block designs in which the responses to contiguous trials are not overlapping in time (see below for a discussion of rapid event related designs). The result is a matrix M ($T \times V$), whose element $m_{t,v}$ is the response estimate at trial t and voxel v. This matrix is partitioned in training and testing matrices ($\mathbf{M}_{train}$ and $\mathbf{M}_{test}$) which are used in the rest of the analysis.

Activation and Discrimination based univariate feature selection.

In previous fMRI applications of pattern recognition methods (Mitchell *et al.,* 2004, Haynes *et al.*, 2005, Mourao-Miranda *et al.*, 2006), univariate feature selection strategies have been suggested for reducing the dimensionality of the multivoxel space (i.e. number of columns in matrix **M**).

Consider a training set defined as:

$$\{\mathbf{m}_i, c_i\} \; i = 1,..., T_{train}; c_i \in \{+1; -1\}; \mathbf{m}_i \in \Re^V \qquad (1)$$

where $\mathbf{m}_i$ is one row of matrix $\mathbf{M}_{\mathbf{train}}$ and represents a trial in the V dimensional space of the voxels, whose class c_i is known (e.g. the two stimulus conditions).

Introducing the hypothesis that interesting patterns consist of voxels that show a significant stimulus-related BOLD response to any of the two conditions compared to baseline levels justifies the reduction of the number of features based on the univariate selection of 'active' voxels. Furthermore, it simplifies the interpretation of the results as the analysis is restricted to voxels showing neurophysiologically understood responses (but see Haynes *et al.*, 2007 and Discussion).

From the values c_i and $m_{i,v}$ ($v = 1,..., V$) we can compute the following scoring functions:

$$S_{A,+1}(v) = \frac{\overline{m_{+1,v}}}{\sqrt{\frac{\sigma^2_{+1,v}}{n_{+1}}}}, S_{A,-1}(v) = \frac{\overline{m_{-1,v}}}{\sqrt{\frac{\sigma^2_{-1,v}}{n_{-1}}}}, \qquad (2)$$

where $\overline{m_{+1,v}}, \sigma^2_{+,v}$ ($\overline{m_{-1,v}}, \sigma^2_{-,v}$) indicate an estimate of mean response and variance calculated over the n_{+1} (n_{-1}) trials of condition $c = +1$ ($c = -1$) at voxel v.

In this work the use of univariate activation based ranking is twofold. First, as a feature selection step to be compared with recursive feature elimination (univAct), in which case we used as ranking criteria the mean between $S_{A,+1}$ and $S_{A,-1}$. We also used activation based ranking as initial univariate feature reduction method in order to select a subset of voxels on which the iterative feature selection procedure is subsequently applied (univActRed). In the latter case we sorted the voxels independently using $S_{A,+1}$ and $S_{A,-1}$ and selected the union of the first V' voxels per condition.

A more restrictive form of univariate feature selection is based on the selection of voxels that show a significant difference between the two conditions (Mitchell *et al.,* 2004, Heynes *et al.*, 2005, Mourao-Miranda *et al.*, 2006). As measures of discrimination ability a parametric (t) or non-parametric (Wilcoxon) statistical test can be used.

From the values c_i and $m_{i,v}$ ($v = 1,..., V$) we can compute the following scoring functions:

$$S_T(v) = \frac{\overline{m_{+1,v}} - \overline{m_{-1,v}}}{\sqrt{\frac{\sigma^2_{+1,v}}{n_{+1}}, + \frac{\sigma^2_{-1,v}}{n_{-1}}}}, \quad (3)$$

$$S_W(v) = \left| \frac{\frac{R_{+1,v}}{n_{+1}}}{n_{-1}} \right| - 1, \quad (4)$$

where $\overline{m_{+1,v}}, \sigma^2_{+,v}$ and $R_{+1,v}(\overline{m_{-1,v}}, \sigma^2_{-,v}, R_{-1,v})$ indicate an estimate of mean response, variance and sum or ranks calculated over the $n_{+1}(n_{-1})$ trials of condition $c = +1(c = -1)$ at voxel v.

The univariate discrimination based selection (univT; univW) is obtained sorting the voxels according to S_T or S_V and selecting the first V' voxels.

It is important to underline the necessity of performing any sort of initial feature selection (activation or discrimination based) only using the training data in order to reduce potential biases in the evaluation of generalization performances.

To better quantify generalization abilities of the univariately selected voxels the scoring functions are computed in cross-validation, i.e. further splitting the training data in different subsets, computing voxel-by-voxel scores on the different sub-splits and then averaging the different scores.

In order to compare univariate feature selection to multivariate feature selection implemented using RFE, we matched the number of voxels selected with the different univariate methods (univT; univW; univAct) to the number of voxels selected by RFE at the different iterations. Furthermore we evaluate the impact of an initial univariate voxel reduction based on activation (univActRed), both on multivariate (RFE) and univariate (univT; univW; univAct) feature selection methods.

Recursive Feature Elimination.

Activation- and discrimination-based feature filtering consider each voxel independently, and thus do not take into account the intrinsic multivariate nature of the fMRI data. Wrapper methods, such as RFE, constitute a multivariate alternative to classical feature filtering that use the classifier itself to discard irrelevant features. Our implementation of RFE can be described with the following pseudocode:

while ($\sim$ stop)

1. Train SVM ($\mathbf{M}_{Train_{RFE_i}}$, $Labels_{RFE_i}$) $i = 1,... N$
2. Compute the scoring function: $S_{RFE}(v) = \frac{\sum_{i=1}^{N} |w_i(v)|}{N}$
3. Sort v based on $S_{RFE}(v)$
4. Eliminate features with smallest scores

end

where $Labels_{RFE_i}$ are the trials' classes in the training set $Train_{RFEi}$, and $w_i(v)$ is the discriminative weight for voxel v as obtained from the SVM training (see below). Before feature selection the SVM is trained multiple times on different subsets of the training data ($Train_{RFEi}$; $i=1,…N$). To perform feature selection the scoring function $S_{RFE}(v)$ is used.

The backward elimination procedure used to search the multidimensional space needs a stopping criterion to be defined. One possible solution is to terminate the algorithm based on the generalization performances (e.g. performance drop compared to previous iteration). A more conservative choice is to proceed from the original feature set to the empty set or a set of desired dimensionality, which in cases of high dimensional feature spaces can be very time consuming (Guyon *et al.*, 2002). When the latter is chosen as stopping criteria the total number of iterations is controlled by the number of voxels discarded at each iteration and the optimal feature set is selected post-hoc based on the highest generalization performances.

Linear Support Vector Machines (binary classification).

Let us consider a training set as in (1). In the general case of overlapping classes (i.e. non-linearly separable classes) the problem of finding the optimal separating hyperplane (defined by the normal **w** and the distance to the origin of the multidimensional space b) that maximizes the distance to the nearest training points of the two classes is defined as:

$$\min_{\mathbf{w}, b, \xi} J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + a \sum_{i=1}^{T_{train}} \xi_i \quad (5)$$

subject to:

$$c_i(\mathbf{w}^T\mathbf{m}_i + b) \geq 1 - \xi_i, i = 1,..., T_{train} \quad (6)$$

and:

$$\xi_i \geq 0, i = 1,..., T_{train} \quad (7)$$

where $\xi_i, i = 1,..., T_{train}$ are slack variables that account for training errors and a is a positive real constant (Suykens *et al.*, 2002). The solution is obtained using Lagrangian methods (Cristianini and Shawe-Taylor, 2000). Classification of new trials $\mathbf{m}_{new}$ is obtained by evaluating:

$$sign\,(\mathbf{w}^T\mathbf{m}_{new} + b). \quad (8)$$

An absolute discriminative map can be obtained considering the vector $|\mathbf{w}|$ (i.e. voxels that contribute the most to the discrimination of the two classes are represented by high values of $|\mathbf{w}|$).

In the present paper, we use a variant of SVM known as ls-SVM. In the classical SVM formulation of Eq. (5-7) the optimal boundary between different classes is obtained by considering only the training point falling on the separating hyperplane (i.e. support vectors). In ls-SVM each training point is weighted in order to obtain the distinguishing hyper-surface (hyper-plane). The optimization problem for the general case of non separable classes is defined as:

$$\min_{\mathbf{w},\,b,\,\xi} J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma \sum_{i=1}^{T_{train}} e_i^2 \quad (9)$$

subject to:

$$c_i(\mathbf{w}^T\mathbf{m}_i + b) = 1 - e_i, i = 1,..., T_{train} \quad (10)$$

where γ is a positive real constant (Suykens *et al.*, 2002).

fMRI data

Simulated time series

We simulated fMRI time series according to a design with two conditions with 30 trials per condition and each trial lasting 14440 *msec* (block design). The functional time series had a simulated TR of 3610 *msec* and functional voxel resolution of 2 x 2 x 2 mm^3. These parameters were used in order to match the experimental design and acquisition parameters used in the real fMRI data set also presented in this paper.

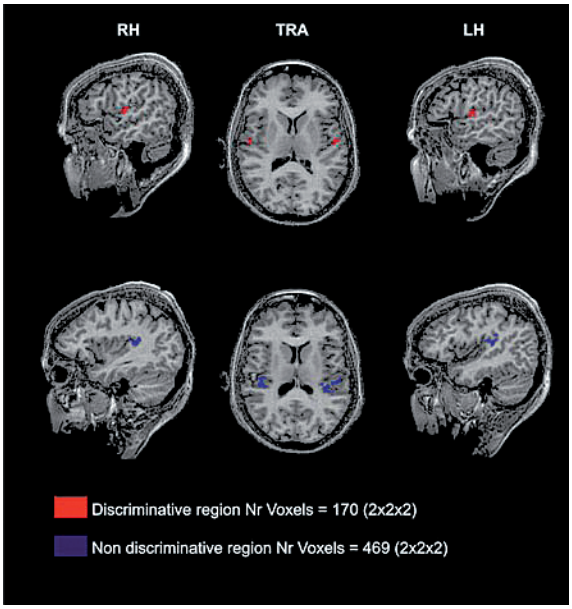

Figure 2: The simulated ROIs projected in the volume of a subject. In red the discriminative ROIs (170 2x2x2 voxels); in blue the active but non discriminative ROIs (469 2x2x2 voxels).

The discriminative voxels (170 in total) were confined to two realistically shaped regions (fig.2 top row) and belonged to one of two populations (condition1>condition2; condition2>condition1) whose spatial distribution was random within the regions. We also simulated neighbouring regions (469 voxels in total)

that responded to both stimulation conditions without carrying specific discriminative information (fig.2 bottom row).

The simulated BOLD responses were obtained by convolving the simulated stimulus with a standard hemodynamic response function modelled using a double gamma function (Friston *et al.*, 1998). At each voxel the simulated activations were added to temporally auto correlated noise obtained as:

$$R_a(t) = \rho_k R_0(t-1) + \sqrt{1-\rho_k}\, R_0(t), \qquad (11)$$

where R_0 is random Gaussian noise and $\rho_k \sim N(0.5, 0.1)$ controls the amount of auto correlation at voxel *k*.

We simulated the data at the level of the original time series and not at the level of the matrix (**M**) as it allows us to examine also the influence of the trial estimation step. For each active voxel we varied the signal-to-noise ratio (i.e. the response amplitude compared to the noise standard deviation; SNR), the contrast-to-noise ratio (i.e. the response differences compared to the noise standard deviation; CNR) and the variability of the BOLD responses to trials of the same stimulus condition (varBOLD), the latter defined in terms of percent of variability compared to the maximum response. Three different levels for SNR [0.3; 0.5; 0.8], CNR [0.1; 0.3; 0.5] and varBOLD [10%; 20%; 60%] were used to produce 27 simulated fMRI data sets.

Simulated functional time series were used to test the performances of a purely multivariate feature selection strategy (RFE). RFE was compared, matching the number of selected voxels, to the performances of SVM based classification preceded by purely univariate feature selection strategies based on t-test or Wilcoxon (univT; univW) and univariate activation based selection (univAct).

Furthermore using univariate activation based ranking as initial voxels reduction strategy (i.e. selecting the union of the 2000 most active voxels for both conditions; univActRed) we evaluated its impact on multivariate feature selection (univActRed+RFE). Matching the number of selected voxels we compared univActRed+RFE to methods in which the same initial activation based reduction was followed by different univariate selection strategies (univActRed+univT; univActRed+univW) and to the case in which only activation based selection was used (univActRed+univAct).

The RFE procedure comprised ten feature selection steps and each step was performed after training the SVM five times (*N=5*). In all the cases the number of discarded voxels at every step was computed so that the size of the final feature set equals the number of simulated discriminative features.

We quantitatively assessed the differences in sensitivity between the various methods using Receiver Operative Characteristics (ROCs) curves computed based on the absolute discriminative maps obtained from each analysis. As a

figure of merit we computed the area under the curve in the false positive rate interval [0, 0.01] (Skudlarski *et al.*, 1999, Sorenson and Wang, 1996, Fadili *et al.*, 2000). The sensitivity of the maps obtained using the different MVPA methods was compared to a conventional univariate analysis (GLM, t-test) in which the entire data sets were used and a design matrix consisting of three predictors (i.e. one for each condition and one accounting for a linear trend) was fitted to the data. The resulting absolute t-maps (one for each simulated data set) were used to compute ROC curves and consequently the area under the curve in the false positive rate interval [0, 0.01].

To evaluate the generalization performances (i.e. accuracies) at different CNR, SNR, varBOLD and feature selection levels we repeated all analysis ten times, each time leaving out a different subset of ten trials for each condition.

Real data

We examined the performances of our approach on real data using a time series from an auditory experiment on sound categorization performed in a 3T system (Siemens Allegra). Functional runs consisted of 23 axial slices obtained with a T2-weighted gradient echo, EPI sequence (TR 3.6 s; FOV 256 x 256; matrix size 128 x 128, voxel size = 2 x 2 x 2 mm^3). Anatomical images were obtained using a high resolution (1 x 1 x 1 mm^3), T1-weighted sequence.

Stimuli consisted of 800 *msec* tonal sounds of four different categories (cats, girls [singing female voices], guitars and tones). The sounds were matched not only in length and RMS power but also in the temporal profile of the fundamental frequency, such that the perceptual pitch could be considered identical across categories. Stimuli were presented in blocks of four during silent periods between TRs, each block lasting 14440 *msec*. Stimulation blocks were followed by blocks of silence lasting 14440 *msec*. Each run consisted of 15 trials per condition presented in a pseudo-random order and lasted 30 *min* approximately. Results presented in this article were obtained using two functional runs of one subject.

The fMRI data sets were subjected to a series of pre-processing operations. (1) Slice-scan-time correction was performed by resampling the time courses with linear interpolation such that all voxels in a given volume represent the signal at the same point in time. (2) Head movements were detected and automatically corrected by minimizing the sum of squares of the voxel-wise intensity differences between each volume and the first volume of the run. Each volume was then resampled in three-dimensional space according to the optimal parameters using trilinear interpolation. (3) Temporal high-pass filtering was performed to remove temporal drifts of a frequency below seven cycles per run. (4) Temporal low pass filtering was performed using a gaussian kernel with FWHM of two data points.

(5) After co-registration to the anatomical images collected in the same session the functional volumes were projected into Talairach space. (6) Moderate spatial smoothing with a Gaussian kernel of FWHM of three millimetres was performed on the volume time series.

After pre-processing, the two functional time series were used for the SVM based analysis as described in figure 1, which produced a total of 30 trials per condition.

In particular we employed a purely multivariate feature selection strategy (RFE). RFE was compared, matching the number of selected voxels, to the performances of SVM based classification preceded by purely univariate feature selection strategies based on t-test or Wilcoxon (univT; univW) and univariate activation based selection (univAct).

Using univariate activation based ranking as initial voxels reduction strategy (i.e. selecting the union of the 2000 most active voxels for all conditions; univActRed) we evaluated its impact on multivariate feature selection (univActRed+RFE). Matching the number of selected voxels we compared univActRed+RFE to methods in which the same initial activation based reduction was followed by different univariate selection strategies (univActRed+univT; univActRed+univW) and to the case in which only activation based selection was used (univActRed+univAct).

For the RFE procedure the same settings as for the analysis of the simulated data were used (ten feature selection levels; *N=5*; percentage of discarded voxels per feature selection step).

Generalization accuracies were estimated by repeating all analysis ten times each time leaving out a different subset of ten trials for each condition.

The same data set was also subjected to conventional univariate statistical analysis using BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). For all six possible contrasts, statistical parametric maps were computed searching for voxels that discriminated between conditions consistently in the two functional runs (conjunction analysis; Nichols *et al.*, 2005) and were thresholded using false discovery rate (FDR, *q=0.05*).

Results

We compared different voxel selection methods in terms of their sensitivity to the true discriminative voxels (ROC analysis; Figure 3a, whole brain analysis; Figure 4a, after initial activation based voxel reduction) and generalization performances (Figure 3b, whole brain analysis; Figure 4b, after initial activation

based voxel reduction). In what follows we detail the results obtained on the simulated and real fMRI data. For comparison, ROC power obtained using conventional univariate statistical parametric mapping (GLM, bold line in Figure 3a and 4a), chance level (bold line in Figure 3b and 4b) and generalization performances obtained using only the simulated discriminative voxels (ROI; dotted line in Figure 3b and 4b) are reported.

Simulated fMRI Data

Whole brain analysis. When starting the analysis from the whole set of voxels (), selecting voxels univariately based on their activation results in higher ROC power (Figure 3a) and higher generalization performances (Figure 3b) than univariate discrimination based voxel selec-

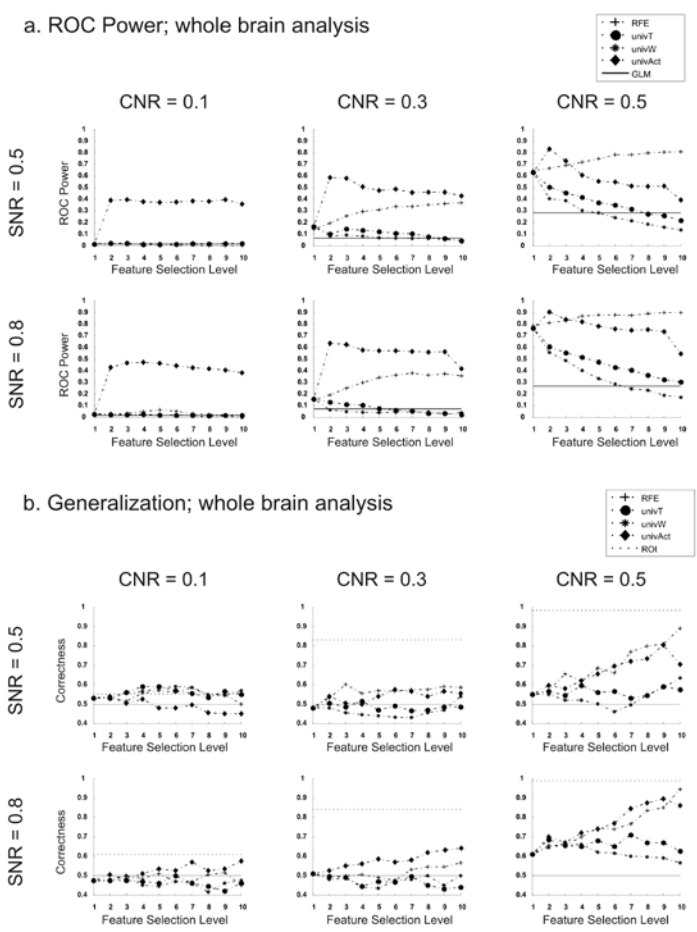

Figure 3: Results obtained on the whole brain analysis using different feature selection strategies at different CNR and SNR value (varBOLD fixed at 10% of the maximum response).

a)ROC power (defined as the area of the ROC curve in the false positive range of [0; 0.01]) of SVM based maps obtained for different numbers of selected voxels starting from the whole brain. b)Generalization performances of SVM based classifier are plotted for different number of selected voxels starting from the whole brain.

Compared feature selection schemes are: 1) RFE; 2) univariate discrimination based selection (univT; univW); 3) univariate activation based slection (univAct). Different methods are compared to classification obtained using only the discriminative voxels (dotted lines). The bold line represents the chance level (0.5).

tion strategy (RFE) iteratively improves sensitivity to the discriminative pattern at CNR = 0.3 and CNR = 0.5 (Figure 3a) and improves generalization at CNR = 0.5 (Figure 3b) but not at very low CNR levels (CNR = 0.1). Compared to RFE, activation based voxel selection has an advantage in terms of ROC power (Figure 3a) at CNR = 0.1 and CNR = 0.3. On the contrary, at the highest CNR level (CNR = 0.5) RFE outperforms activation based voxel selection both in terms of sensitivity (Figure 3a) and generalization (Figure 3b). These results can be explained considering the nature of the simulated discriminative voxels, which are few compared to the entire set and present activation levels (depending on the SNR) above the baseline noise. At high CNR levels, the discriminative information is sufficient to drive the multivariate search despite the large number of irrelevant voxels. At lower CNR levels this is not the case and selecting univariately

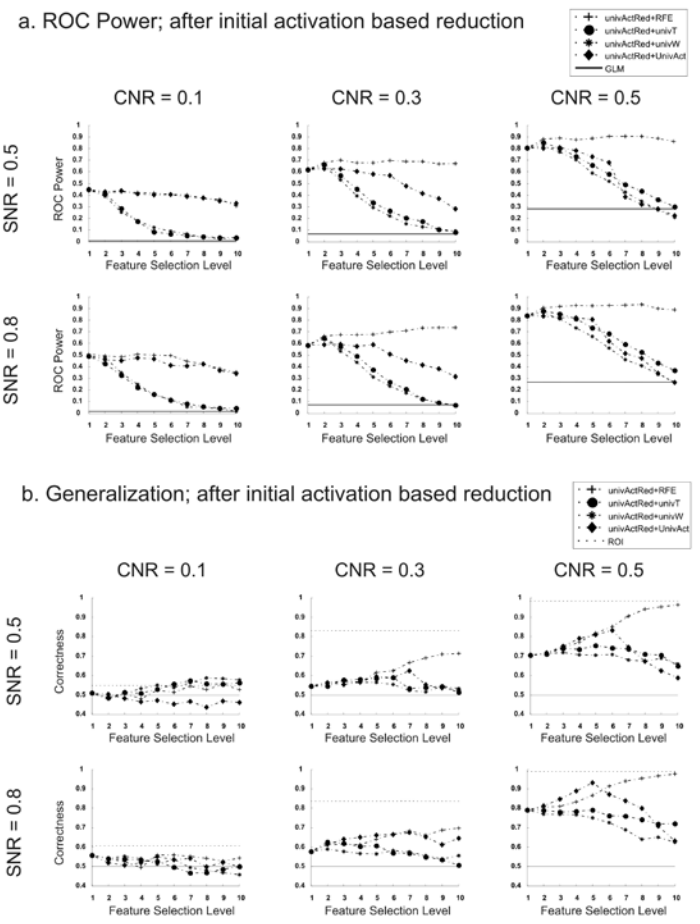

Figure 4: Results obtained from the combination of an initial univariate activation based voxel reduction and different voxel selection strategies. Results are reported at different CNR and SNR value (varBOLD fixed at 10% of the maximum response).

a)ROC power (defined as the area of the ROC curve in the false positive range of [0; 0.01]) of SVM based maps obtained for different numbers of selected voxels starting from a subset of voxels selected using activation based reduction (univActRed).
b)Generalization performances of SVM based classifier are plotted for different number of selected voxels starting from a subset selected using univariate activation based reduction (univActRed).

Compared feature selection schemes are: 1) RFE; 2) univariate discrimination based selection (univT; univW); 3) univariate activation based slection (univAct). Different methods are compared to classification obtained using only the discriminative voxels (dotted lines). The bold line represents the chance level (0.5).

the voxels based on their activation proves to be more effective. This suggests the combination of univariate activation based voxel selection and RFE as a promising strategy for MVPA.

Combination of univariate and multivariate voxel selection. Figure 4 shows that – after univariate activation based voxel reduction – iteratively pruning the voxels based on their multivariate information (univActRed+RFE) clearly outperforms both in terms of sensitivity (Figure 4a) and generalization performances (Figure 4b) all other feature selection strategies. This same strategy provides the highest performances also compared to whole brain analysis (compare figure 4 and 3) and allows approaching close-to-optimal levels of classification, as defined by those obtained using only the discriminative voxels (dotted lines in figure 4b). Improvements are not observed only at CNR = 0.1, this can be due to the fact that at this CNR level, even using only the discriminative voxels, classification performances are close to chance level.

Performances decrease when the variability of the BOLD response increases. Figure 5 shows for univActRed+RFE ROC power and generalization at different SNR and varBOLD levels for different CNRs. Both sensitivity and classification performances are negatively affected by the variability of the BOLD response. As expected, the decrease in performances with increasing variability is stronger at lower CNRs and SNRs where the intra-class distance in the multivariate space is lower.

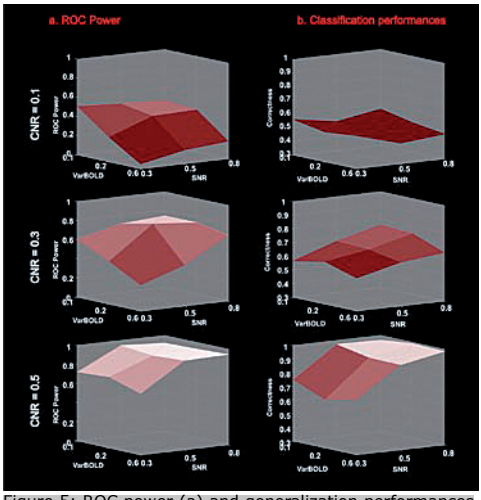

Figure 5: ROC power (a) and generalization performances (b) obtained using univActRed+RFE at different SNRs and varBOLD levels reported for different CNR values.

Real Data

Figure 6 and 7 show the results obtained using SVM based classification and different feature selection schemes on the real fMRI data.

RFE improves Single Trials classification performances. Figure 6 shows the SVM based classification performances for the six possible contrasts at different feature selection levels for different feature selection methods. Both when starting from the whole brain (figure 6a) or after an initial voxel reduction based on single conditions activation levels (figure 6b), the highest classification performances for each contrast are obtained using RFE. Im-

provements in classification of single trials are visible for each contrast especially when RFE follows an initial univariate feature selection based on activation measures (figure 6b). In Table 1 we report for each binary classification the percentage of correct classification obtained using univActRed+RFE, the optimal

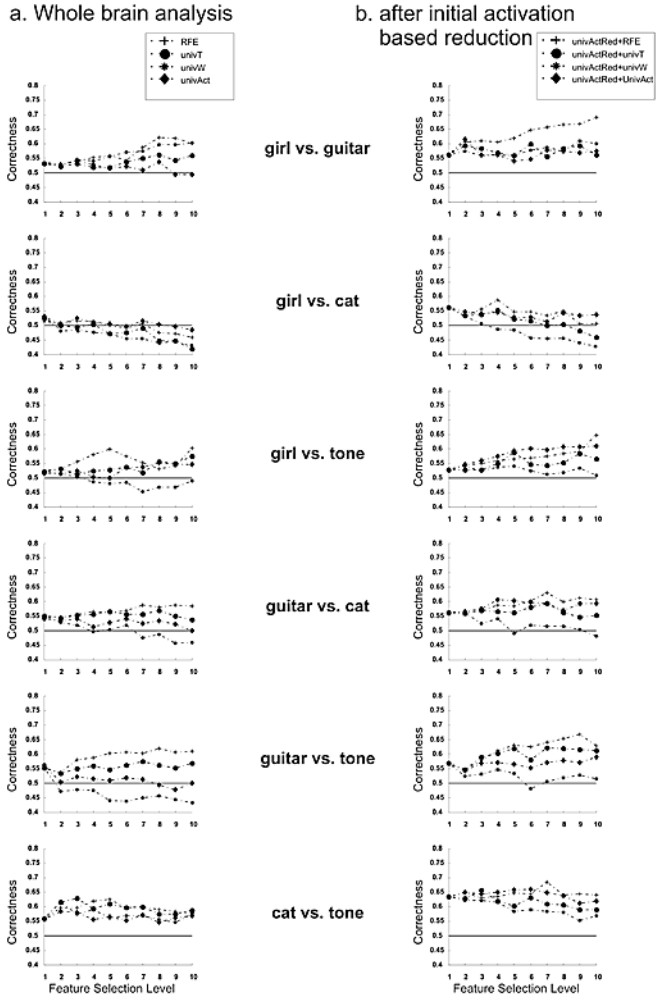

Figure 6: Classification performances obtained on the real data set for each binary comparison in the discrimination of sound categories; a)Performances of different feature selection schemes on whole brain analysis; b)Performances of different feature selection schemes after an initial univariate activation based voxel reduction.

feature selection level (and the corresponding number of voxels used in the classification), the size of the improvement (highest classification performance minus classification performance at the initial feature selection step). For comparison we also report the percent difference in classification between univActRed+RFE

and the closest performing method at the same feature selection level.

Table 1

	Girl/Guitar	Girl/Cat	Girl/Tone	Guitar/Cat	Guitar/Tone	Cat/Tone
Correct Classification	66%	58%	65%	61%	65%	67%
Optimal Level (univActRed+RFE)	10 (NrVox = 236)	4 (NVox = 2002)	10 (NrVox = 236)	7 (NrVox = 687)	9 (NrVox = 337)	7 (NrVox = 687)
Improvement Size (univActRed+RFE)	10%	2%	12%	6%	9%	3%
Difference to closest method	7% (univActRed + univW)	3% (univActRed+ univT)	5% (univActRed + univAct)	2% (univActRed + univAct)	5% (univActRed+ univT)	3% (univActRed + univAct)

Summary of results obtained on the real data analysis when RFE is applied after initial voxel reduction obtained using univariate activation based ranking.
Improvement size is defined as the highest generalization performances minus the generalization performances at the first feature selection level.
Optimal feature selection level (number of voxels used in the classification are reported in brackets) is defined as the level at which the highest classification performances are obtained, for each contrast.
The difference to the closest performing method (reported in brackets) at the same feature selection level is also reported.

Figure 7 shows detailed results obtained using classical univariate mapping (GLM, first column) and RFE after univariate activation based voxel reduction (univActRed+RFE) (second column) for the six different discriminations. GLM based contrast maps show voxels with significant activation differences that were consistent in the two functional runs and were thresholded using false discovery rate (FDR, q=0.05), univActRed+RFE absolute discrimination maps represent the best 20% of selected voxels at the optimal feature selection level, as defined by the highest generalization performances. Generalization performances at different feature selection levels are reported as median, lower and upper quartile across the different iterations. The optimal feature selection level is highlighted and chance level is reported for comparison as a dashed line.

The sixth contrast (cat vs. tone) shows significant bilateral univariate differences. The same areas are highlighted as most discriminative using univActRed+RFE, with highest generalization performances (0.67) obtained at the seventh iteration. Similar generalization performances are obtained for the other contrasts (girl vs. guitar: 0.66; girl vs. cat: 0.58; girl vs. tone: 0.65; guitar vs. cat: 0.61; guitar vs. tone: 0.65) at different feature selection levels. Note that none

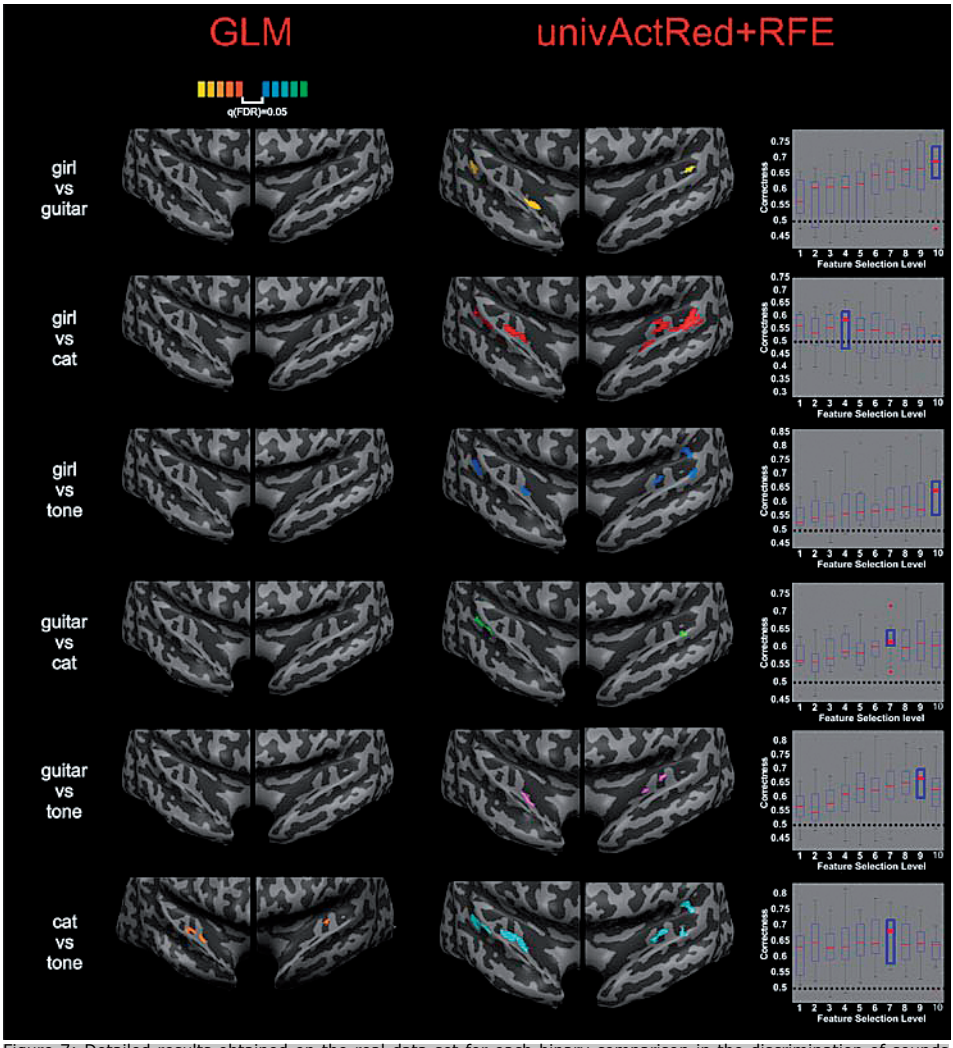

Figure 7: Detailed results obtained on the real data set for each binary comparison in the discrimination of sounds categories. GLM maps (first column) are projected over the inflated cortex of the subject and thresholded at a q=0.05 (FDR corrected). Discriminative maps obtained using univActRed+RFE at the optimal feature selection level (second column) are projected over the inflated cortex of the subject and thresholded in order to visualize the best (most discriminative) 20% of the voxels. Generalization results (median, lower, upper quartile and dispersion) (third column) obtained using univActRed+RFE at different feature selection levels, the best level is highlighted.

of them show significant univariate differences. For all contrasts, maps obtained with univActRed+RFE highlight bilateral discriminative regions along the superior temporal gyrus, both anterior and posterior to the primary auditory regions located along the Heschl's gyrus.

In order to compare the real data results with the simulation results we computed the SNR, CNR and BOLD variability of the real data for each contrast in the voxels that produced the highest generalization results. Using these computed

values and the generalization performances we determined the closest simulation case and estimated from the corresponding parameter values the ROC power for the real data. The results are reported in Table 2. Accepting a false positive rate in the interval [0, 0.01] all contrasts are in the ROC Power interval [0.61, 0.67].

Table 2

	Girl/Guitar	Girl/Cat	Girl/Tone	Guitar/Cat	Guitar/Tone	Cat/Tone
SNR	0.26	0.28	0.25	0.30	0.25	0.26
CNR	0.32	0.26	0.30	0.31	0.30	0.34
varBOLD	0.11	0.12	0.12	0.11	0.12	0.11
%Correct	66	58	65	61	65	67
ROC Power	0.65	0.61	0.66	0.67	0.66	0.67

Generalization performances on the real data set reported together with estimated CNR, SNR, BOLD variability and sensitivity (i.e. ROC power for the false positive rate interval [0; 0.01]).

Discussion

Differently from conventional univariate statistical analyses, machine learning techniques take advantage of the multivariate nature of the fMRI data and highlight maximally discriminative spatial patterns. While these methods offer a sensible advantage compared to conventional univariate mapping in the case of low contrats-to-noise scenarios, the main challenge in their application to fMRI is dealing with the large number of voxels in combination with a rather low number of trials of a typical scan. Performances of pattern recognition methods such as SVMs are in fact known to degrade with the increasing number of irrelevant features.

In the present article we have described and evaluated an approach for fMRI pattern discrimination analysis based on Support Vector Machines and a combination of univariate and multivariate feature selection strategies. Using this approach, the search for multivoxel discriminative patterns is iterative and data driven, thus minimizing the number of required spatial assumptions on the location and extent of the patterns.

Compared to previous approaches employing whole brain analyses (Mourao-Moranda *et al.*, 2005), the evaluated method increases the sensitivity for the

discriminative patterns, especially when they include a relatively small number of voxels compared to the whole data set (sparse discriminative patterns). This method can thus be seen as a useful solution when specific hypotheses on the localization (Haynes *et al.*, 2005, Kamitani *et al.*, 2005) and/or dimension (Kriegeskorte *et al.*, 2005) of the spatial patterns are not available.

In our approach, the search of patterns is based on the Recursive Features Elimination (RFE) algorithm (Guyon *et al.*, 2002), which iteratively eliminates the least discriminative features based on multivariate information as detected by the classifier (Support Vector Machine) itself.

In a previous publication Carlson *et al.* (2003) used a "knock out" procedure to examine the degree of overlap in information between the representations of different object categories (chairs, faces and houses) in the visual cortex. In particular the procedure aimed to compare the reduction in classification performances of a stimulus category (e.g. chairs) in two cases: first, when the discriminant direction between a stimulus category and all other categories was removed (chairs vs. faces and houses) and second, when the discriminant direction of another stimulus was removed (faces vs. chairs and houses). This "virtual lesion" approach was implemented projecting the multivoxel patterns on the different category-specific discriminant hyperplanes (i.e. removing the direction of the category specific maximum discrimination in the multivoxel space) and subsequently evaluating the performance losses for each category. The authors showed that removing a category specific discriminant reduced the classification of all other categories only in part. These results suggest that there is not a complete overlap between the representations of the different object categories in the visual cortex. While the aim of the knock out procedure of Carlson *et al.* is to evaluate the similarity between the multi voxel patterns elicited by different stimulation conditions, RFE aims to optimize in a multivariate and data driven way the discriminative information between different categories. In particular, while RFE removes at each iteration the least discriminant voxels the knock out procedure of Carlson *et al*. removes a direction in the feature space which is a weighted average of all voxels.

Results of our simulations show that the combination of RFE and univariate activation based reduction of voxels ensures the highest sensitivity and generalization performances (see figure 4). In particular, when RFE is applied after an initial univariate activation-based voxel reduction there is a sensible advantage compared to the case the same initial voxel reduction is followed by univariate discrimination or activation based selection, especially at very low CNRs (see figure 4 a-b). This is a consequence of assuming that, at single voxel level, BOLD changes of a condition compared to the baseline are greater than BOLD differ-

ences between conditions (SNR > CNR), which appears as a realistic assumption in most fMRI studies. This result confirms previous comments on the use of univariate feature selection methods to MVPA (Mitchell *et al.,* 2004, Mourao-Miranda *et al.*, 2006). Because of the reduced sensitivity of univariate statistics at low SNRs/CNRs the most appropriate choice for combining univariate and multivariate features selection is to use rather liberal thresholding for univariate selection (which prevents the exclusion of potentially informative voxels) and further discard irrelevant voxels based on the multivariate scoring function.

Also alone, the recursive proves to be more sensitive than conventional univariate analysis (General Linear Model; figure 3a) and shows sensible improvements with decreasing number of voxels (figure 3 a-b). In our simulations, indeed, both ROC power and generalization performance increased with feature selection level, with the latter approaching optimal values, i.e. those obtained using only the simulated discriminative voxels (figure 4b). Note that the superiority of the combined approach (univActRed+RFE) compared to the purely multivariate approach (RFE) is due to the chosen strategy to use a constant value for the total number of feature selection steps in the different methods. Better performances might be obtained allowing the multivariate approach to exhaustively search the whole set of features with a larger number of smaller steps. However, this would require a much longer computational time.

Our simulations did not consider the case in which discriminative patterns are not represented by regions that do not show a global main effect of activation (Haynes *et al.*, 2007), in which case using RFE without univariate activation based pre selection may prove to be more sensitive. More generally, available a-priori information on the nature of the effects of interest (e.g. presence of a global main effect) or on its location may aid and guide the chosen feature selection strategy. As shown by the simulation results, perfect anatomical knowledge of the location of the discriminative patterns (ROI approach) proves to be the most sensitive method. Our RFE-based approach, on the other hand, allows searching for discriminative patterns in a more “data driven” way with no initial assumption on their location. The analysis of the real data shows an illustrative case. For sound categorization, little is known on the exact localization of the discriminating patterns within the human auditory cortex and defining anatomical or functional landmarks is not straightforward. Using RFE we were able to map these discriminative patterns and improve the generalization performance compared to the initial anatomical selection of voxels.

Furthermore our simulations were limited to the case of two sparse and spatially distributed populations of voxels without explicit covariances between them (e.g. functional and effective connectivity). The observed superiority of the

multivariate analysis compared to the mass-univariate GLM is thus due to the integration of weak univariate differences irrespective of the sign of the discriminative information, which also explains its advantage compared to conventional smoothing (Kriegeskorte *et al.*, 2006). In such cases, combining RFE with classifiers other than the ls-SVM (e.g. GNB) would produce similar results. More generally, however, the presence of functional and effective connectivity among voxels may affect the final outcome of our method. In fact, while the proposed RFE procedure can be applied to any algorithm, the weighting of individual features is algorithm-dependent and may be influenced by the way the classifier handles the covariance between the features.

We tested the same approaches on fMRI time series obtained in an auditory experiment with sounds from different categories. In line with the results from the simulations, RFE preceded by activation based univariate voxel reduction selection (univActRed+RFE) produced the highest generalization performances and the recursive approach to feature elimination improved generalization performances in all contrasts (figure 6; Table 1). Classical univariate mapping failed to reveal discriminative regions for all contrasts except the sixth (cat vs. tone) (figure 7, first column). As expected, in this latter case, discriminative maps produced by univActRed+RFE overlapped with GLM contrast map and were accompanied by above chance generalization performances (cat vs. tone: 0.67). In all other contrasts, despite the lack of statistically univariately significant voxels in standard GLM analysis, our approach reached comparable generalization performances (girl vs. cat: 0.58; girl vs. guitar: 0.66; girl vs. tone: 0.65; guitar vs. cat: 0.61; guitar vs. tone: 0.65). The discriminative spatial patterns as highlighted by univActRed+RFE comprise multiple non neighbouring regions in the anterior and posterior portions of the superior temporal gyrus, in both the right and left auditory cortex. These results are consistent with the notion of a 'what' auditory processing stream originating in the superior temporal areas, anterior to the Heschl's gyrus (Belin *et al.*, 2000a, Belin *et al.*, 2000b, Rauschecker *et al.*, 2000, Lewis et al., 2005) and with recent fMRI studies which point to a relevant role of STS in the representation and processing of complex sounds (Warren *et al.*, 2005). A full account of these results, including a group, is given in Staeren et al. (submitted).

One possible drawback of the application of RFE is the backward elimination strategy, which requires setting of several parameters, the most relevant being the number of iterations and the number of features to discard at each iteration. Searching exhaustively the whole feature set would require a large number of iteration with few discarded voxels at each iteration. Especially in fMRI, however, such an approach would result in very long computational time. In the present paper we selected a relatively small but practically feasible number of feature

selection steps (10) and discarded, at each iteration, a fixed proportion of the current number of features computed based on the desired final set size. While arbitrary, this choice proved to be effective both in the case of the simulations and of the real data. It should be noted, however, that different data sets may require different settings. Other multivariate feature selection methods, such as the embedded algorithm by Rakotomamonjy (2003) suffer from similar problems, as the discriminative feature set size has to be chosen after the optimization is terminated. The application of these methods, thus, requires heuristic choices, compromising practical feasibility and optimal search criteria.

The single trial estimation procedure we outlined in this paper was designed for block or slow event related designs for which one may derive a response-pattern estimate for each block/event (or even TR). This is not possible with rapid event related designs. In these cases, applying MVPA requires subdividing the measurements in many sub parts, each one including a sequence of trials that allows for estimating condition response patterns. Assuming linearity of BOLD responses, one obtains a response-pattern estimate from each of these sub-runs that can then be used as those obtained from a slow event related design. After trial estimation, application of RFE is identical as in blocked or slow event related designs.

Another methodological consideration regards the possibility of defining a statistical threshold for the maps produced by the SVM classifier. As described in previous publications a statistical threshold for the single subject maps could be obtained by permutation testing (Mourao-Miranda *et al.*, 2005, Wang *et al.*, 2007) or randomization (Kriegeskorte *et al.*, 2006). Another possibility for the group level maps is to perform random effect analysis across subjects (Wang *et al.*, 2007).

The ROC analysis performed to evaluate the sensitivity of our approach on the simulated data sets is independent of the threshold as the true positive and false positive ratios are computed for a range of thresholds. For real data, maps are thresholded such that 20% of the most discriminative voxels at the optimal feature selection level are shown. Note that in this case the selected voxels provide significant generalization performances in the classification of new trials and the results reported in Table 2 indicate that the proposed approach is sensitive to the underlying discriminative patterns.

Conclusions

We illustrated different strategies to perform feature selection for pattern discrimination analysis of fMRI data and introduced a novel, data-driven feature selection strategy that uses multivariate information. Our results show that the combination of univariate (activation based) and multivariate feature selection outperforms other techniques when no a priori information is available on the size and location of the pattern of interest.

The proposed method could be extended to the multivariate analysis of data from other imaging modalities (perfusion MRI, PET, optical imaging, EEG, MEG) or to their combination. In the latter, feature elimination could be applied in the multidimensional space of features extracted from different methods (e.g. voxels of fMRI and time points of EEG) and used to reveal the most discriminative set of multi modal features.

Acknowledgments

Financial support from NWO (MaGW-VIDI grant 452-04-330) to EF is gratefully acknowledged.

References

Belin, P. and Zatorre, R. J. (2000a). 'What', 'where' and 'how' in auditory cortex. Nat. Neurosci.; 3(10): 965-966.

Belin, P., R. J. Zatorre, Lafaille, P., Ahad, P., Pike, B. (2000b). Voice-selective areas in human auditory cortex. Nature; 403(6767): 309-312.

Cox, D., Savoy, R. 2003. Functional magnetic resonance (fMRI) "Brain Reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage; 19(2): 261-270.

Cristianini and Shawe-Taylor 2000. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.

Fadili, M.J., Ruan, S., Bloyet, D., Mazoyer, B. (2000). A multi-step unsupervised fuzzy clustering analysis of fMRI time series. Hum. Brain Mapp.; 10:160–178.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., Turner, R. (1998). Event-Related fMRI: Characterizing Differential Responses. Neuroimage; 7(1): 30-40.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning; 46: 389-422.

Guyon, I., Elisseef, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research; 3: 1157-1182.

Hanson, S.J., Halchenko Y. (in press). Brain reading using full brain support vector machines for object recognition: there is no face identification area. To appear in Neural Computation.

Haynes JD, Rees G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci.; 8(5):686-91.

Haynes, J.D., Rees, G. (2006). Decoding mental states from brain activity in humans. Nat. Rev. Neurosci.; 7(7):523-34.

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E. (2007). Reading hidden intentions in the human brain. Current Biology; 17: 323-328.

Kamitani Y, Tong F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci.; 8(5):679-85.

Kohavi, R., John, G. (1997). Wrappers for feature selection. Artificial Intelligence; 97(1-2): 273-324.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America; 103(10): 3863-3868.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. Neuroimage; 26(2): 317-329.

Lewis, J. W., J. A. Brefczynski, Phinney, R. E., Janik, J. J., DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. Journal of Neuroscience; 25(21): 5148-5158.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X. (2004). Learning to decode cognitive states from brain images. Machine Learning; 57: 145-175.

Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage; 28(4):980-95.

Mourao-Miranda J, Reynaud E, McGlone F, Calvert G, Brammer M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. Neuroimage; 33(4):1055-65.

Nichols TE, Brett M, Andersson J, Wager T and Poline J-B (2005). Valid conjunction inference with the minimum statistic. NeuroImage, 25:653–660.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. Sep; 10(9):424-30.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. Journal of Machine Learning Research; 3: 1357-1370.

Rauschecker, J. P., Tian, B. (2000). Mechanisms and streams for processing of „what“ and „where“ in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America; 97(22): 11800-11806.

Skudlarski, P., Constable, R.T., Gore, J.C. (1999). ROC analysis of statistical methods used in fMRI: individual subjects. Neuroimage; 9: 311–329.

Sorenson, J.A., Wang, X. (1996). ROC method for evaluation of fMRI Techniques. Magn. Reson. Med.; 36: 737–744.

Suykens, J.A.K., Van Gestel, T., De Barbanter, J., De Moor, B., and Vanderwalle, J., 2002. Least Squares Support Vector Machines. World Scientific Publishing.

Wang, Z, Childress, A.R., Wang, J, Detre, J.A. (2007). Support vector machine

learning-based fMRI data group analysis. Neuroimage; doi: 10.1016/j.neuroimage.2007.03.072.

Warren, J. D., Jennings, A. R., Griffith, T.D. (2005). Analysis of the spectral envelope of sounds by the human brain. Neuroimage; 24: 1052–1057.

Who's saying what? Decoding speech content and speaker identity from auditory cortical activity. 4

Abstract
We decipher speech content ('what' is being said) and speaker identity ('who' is saying it) from observations of sound-evoked brain activations of listeners. By combining machine learning with functional MRI, we unravel the distributed and overlapping auditory cortical fingerprints of linguistic units (vowels) and speakers' voices. These fingerprints are insensitive to acoustic variations of the input and allow robust brain-based speech decoding and speaker identification.

Based on:
Formisano, E., De Martino, F., Bonte, M., Goebel, R. (2008). Who's saying what? Decoding speech content and speaker identity from auditory cortical activity. (submitted)

In everyday life, we automatically and effortlessly decode speech into language independently of 'who' speaks. Similarly, we recognize a speaker's voice independently of 'what' she/he says. Cognitive and connectionist models postulate that this efficiency depends on the crucial ability of our speech perception and speaker identification systems to extract relevant features from the sensory input and form computationally efficient abstract representations (Luce *et al.*, 2000; McClelland and Elman, 1986; Norris, 1994). A defining property of these representations is their *invariance* with respect to changes of the acoustic input, which ensures efficient processing and confers to both systems a high robustness to noise or to signal distortion.

As a noticeable example of these representations, relevant psycholinguist models consider abstract entities such as phonemes as the building blocks of the chain of computations that transform an acoustic waveform into a meaningful concept (Luce *et al.*, 2000; McClelland and Elman, 1986; Norris, 1994). Phonemes are not contingent to a specific acoustic implementation so that different waveforms can be associated with the same phonemic token (e.g. the same vowel spoken by different persons). Along the same line, there is psychoacoustic evidence that the identification of a speaker relies on the extraction of *invariant* paralinguistic features of his/her voice, such as fundamental frequency and timbre, which do not depend on the actual speech content (Belin *et al.*, 2004).

The neurological basis of speech (phoneme) and voice processing has been investigated in numerous functional MRI (Belin *et al.*, 2000; Binder *et al.*, 2000; Desai *et al.*, 2008; Liebenthal *et al.*, 2005; Obleser *et al.*, 2006; Obleser *et al.*, 2007; Scott *et al.*, 2000), electro- and magneto-encephalography studies (Näätänen, 2001; Näätänen *et al.*, 1997; Obleser *et al.*, 2004; Shestakova *et al.*, 2004). By comparing speech or vocal sounds to control sounds, several cortical regions in the superior temporal cortex have been characterized in terms of their 'selectivity' or 'specialization' for individual features that are relevant to the 'speechness' or 'voiceness' of the stimuli. In particular, results suggest the involvement of a left lateralized cortical network of superior temporal areas (Scott and Johnsrude, 2003) for the processing of speech (phonemes), and of a right lateralized set of auditory regions along the superior temporal sulcus (STS) for the processing of voice and speaker identity (Belin *et al.*, 2004). The timing of these processes has been estimated between 100 to 350 ms after speech/voice onset (Belin *et al.*, 2004; Näätänen, 2001).

The subtraction-based experimental logic used in these previous studies, however, only allows partial and indirect inferences on the nature and properties of the cortical representations of a speech sound and on the neurocomputational mechanisms that our brain uses for transforming the acoustic input into speech

content and/or speaker identity. It remains unknown, for example, whether an abstract (i.e. not contingent to a specific acoustic implementation) representation of speech content (phoneme) and of speaker identity is formed already at level of the auditory cortex or only at following stages of cortical processing.

Here we combine single-trial fMRI with machine learning to directly map the representations of basic linguistic units (vowels) and speakers' voices in the human auditory cortex. Using a novel iterative algorithm that allows modelling of spatially distributed as well as localized (clustered) response patterns (De Martino *et al.*, in review), we asked whether the observation of the neural activation 'fingerprint' of a speech sound (vowel) is sufficient to decipher its content ('what' is being said) and the identity of the speaker ('who' is saying it). Furthermore, we directly test the invariance of these cortical fingerprints with respect to acoustic variations of the input, by examining the performance of a brain based classifier in recognizing speech content from novel speakers and in recognizing speakers with novel speech content.

High spatial resolution (1.5 mm x 1.5 mm x 2 mm) functional images of the auditory cortex were collected while participants listened to speech sounds consisting of 3 Dutch vowels (/a/, /i/, /u/) recorded from 3 native Dutch speakers

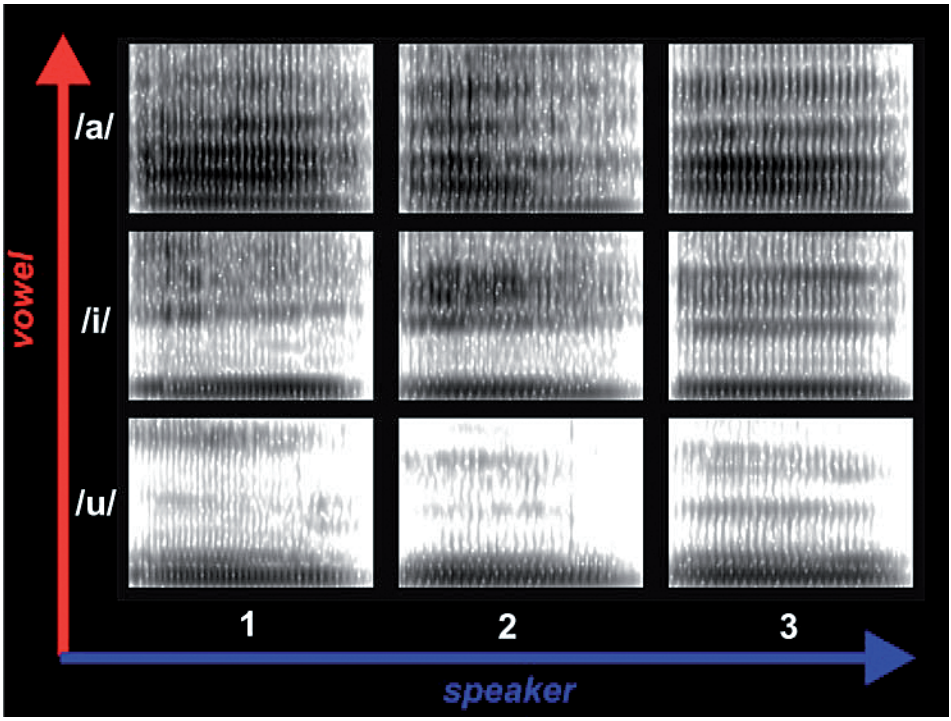

Figure 1: Experimental design and spectrograms of the nine stimuli (3 vowels x 3 speakers).

(sp1, female; sp2, male; sp3 male) (see Methods and Figure 1) and presented in a slow event-related fMRI design.

All sounds evoked significant blood oxygenation level dependent (BOLD) responses in a wide expanse of the superior temporal cortex, including early auditory areas (Heschl's gyrus) and multiple regions in the planum temporale (PT), along the superior temporal gyrus (STG), the superior temporal sulcus (STS) and the middle temporal gyrus (MTG). With univariate contrasts, only weak (below significance) or no response differences between conditions were found (see

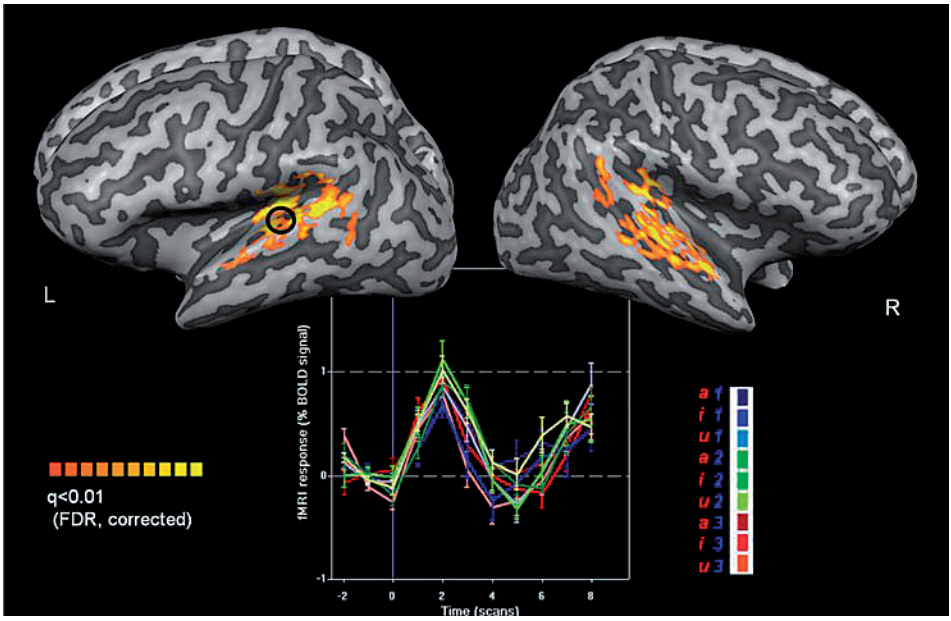

Figure 2: Upper panel: Example of auditory cortical activation in response to the speech sounds as estimated using univariate general linear model analysis (F map, single subject). Lower panel: Event-related BOLD responses to the nine experimental conditions.

Figure 2), did allow decoding of neither the sound content nor the speaker. After this initial analysis, we examined whether jointly considering the activations measured at many cortical locations would make this decoding possible.

We performed two complementary analyses: First, we labelled the stimuli and corresponding response patterns according to the *'vowel'* dimension irrespective of the *'speaker'* dimension, which led to the grouping of stimuli and responses in the three conditions /a/, /i/, and /u/ (*vowel learning*). We then examined whether our recursive machine learning algorithm (see Methods), after being trained with a subset of labelled brain responses (40 trials), would accurately classify remaining unlabelled responses (10 trials, see Methods). In all subjects

and in all possible pair wise comparisons, the recursive algorithm was successful in learning the functional relation between sounds and corresponding spatial patterns and classified the unlabelled sound-evoked patterns accurately (/a/ vs /i/ = 0.65 (mean accuracy), $P = 6 \cdot 10^{-5}$; /a/ vs /u/ = 0.69, $P = 2 \cdot 10^{-5}$; /i/ vs /u/ = 0.63, $P = 4 \cdot 10^{-4}$; see Figure 3c for median and range of these accuracy levels).

Second, we labelled stimuli and corresponding response patterns according to the *'speaker'* dimension irrespective of the *'vowel'* dimension, which led to grouping of stimuli and responses in the three conditions sp1, sp2, and sp3

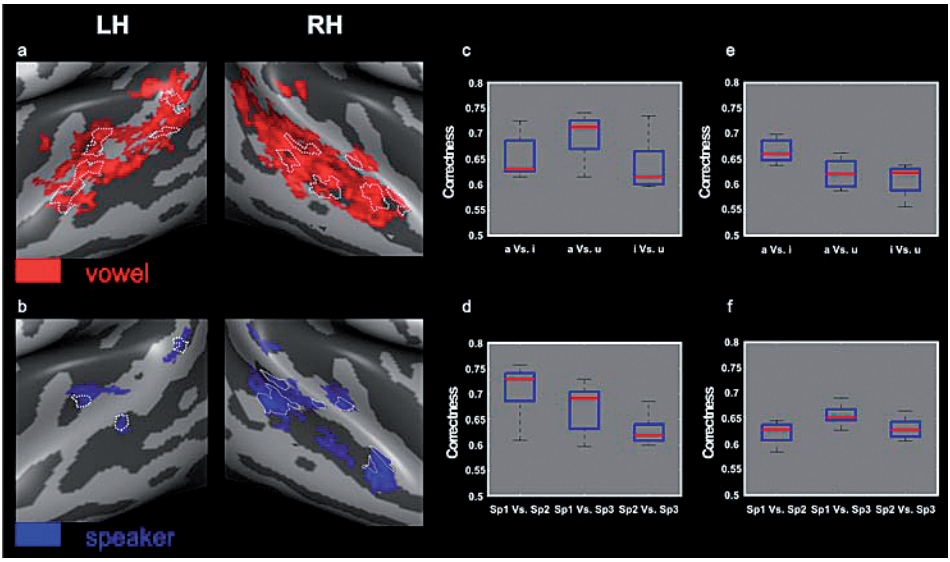

Figure 3: Group discriminative maps (a, b), classification (c, d) and generalization (e, f) accuracies for the all pair wise discriminations of vowels (a, c, e) or speakers (b, d, f) over all subjects (n=7). Group discriminative maps were obtained from cortex-based realignment of individual discriminative maps and represent the cortical locations that are maximally informative with respect to the discrimination of vowels (red color, a) or speakers (blue color, b). A colored vertex in the map indicates that the colored location was present in at least four out of the seven individual maps. Classification accuracies in c and d refer to the brain based machine learning of vowel (speaker) performed with training and testing samples from all speakers (vowels). Generalization accuracies in e and f refer to the brain based machine learning of vowel (speaker) using testing samples from speakers (vowels) other than those used for training. Outlined regions in a, and b, indicate cortical regions that were included in the group discriminative maps for this latter generalization analysis.

(*speaker learning*). Also in this case, machine learning was successful and in all subjects and in all possible pair wise comparisons the classification was accurate (sp1 vs sp2 = 0.70, $P = 3 \cdot 10^{-5}$; sp1 vs sp3 = 0.67, $P = 8 \cdot 10^{-5}$; sp2 vs sp3 = 0.62, $P = 2 \cdot 10^{-5}$; see Figure 3d).

These results indicate that auditory fMRI activation patterns evoked by a presentation of a single sound possess enough information for identifying accurately speech content or identity of the speaker.

To investigate the spatial layout and the consistency across subjects of the

spatial patterns that make this decoding possible, we generated *group discriminative maps*, i.e. maps of the locations that contribute most to the discrimination of conditions (Figure 3a and Figure 3b). These maps were obtained by cortical realignment of individual discriminative maps, which included only locations that "survived" the recursive elimination of irrelevant features in the algorithm (see Methods). A coloured vertex in the group maps indicates that the corresponding location was present in at least four out of the seven individual discriminative maps.

The activation patterns that discriminated between the three *vowels* (red maps in Figure 3 and Figure 4) were widely distributed in the superior temporal cortex. Discriminative patterns included regions in the anterior-lateral portion of the Heschl's gyrus/Heschl's sulcus, in the PT (mainly in the left hemisphere), and extended portions the STS/STG (both hemispheres). The activation patterns that discriminated between the three *speakers* (blue maps in Figure 3 and Figure 4) were more localized than those obtained for vowel discrimination. These patterns included regions in the lateral portion of the Heschl's gyrus/Heschl's sulcus (both hemispheres), located in the posterior adjacency of similar regions described for the *vowel* discrimination (see the superimposition of maps in Figure 4) and three prominent and clustered regions along the anterior-poster axis of the right STS.

The group discriminative maps for *'vowel'* and *'speaker'* learning highlighted three noticeable differences between the cortical representations of speech content and speaker identity. First, the overall distribution of informative locations across hemispheres was symmetrical for the *'vowel'* but highly asymmetrical - with right hemispheric dominance - for the *'speaker'* discriminative map. Second, whereas the maps resulting from the *'speaker'* learning consisted of a small set of clustered regions, the *'vowel'* learning analysis resulted in extended patterns covering most of the activated auditory cortex. Third, although some overlap exists between the two discriminative maps (purple colour in Figure 4), the cortical representations of *'speaker'* and *'vowel'* are clearly disjoint (see red and blue colour in Figure 4).

Encouraged by the results of these analyses, we tested the capability of our machine learning algorithm to decipher the brain activity into speech content and speaker identity also in the case of completely novel speech stimuli (i.e. stimuli not used during the training). We thus trained the iterative learning algorithm in discriminating vowels with samples from one speaker (e.g. /a/ vs /i/ for sp1) and tested the accuracy of this discrimination in the other speakers (e.g. sp2 and sp3). Analogously, we trained the learning algorithm in discriminating speakers with samples from one vowel (e.g. sp1 vs sp2 for /a/) and tested the accuracy of this discrimination in the other vowels (e.g. /i/ and /u/). Note that with this train-

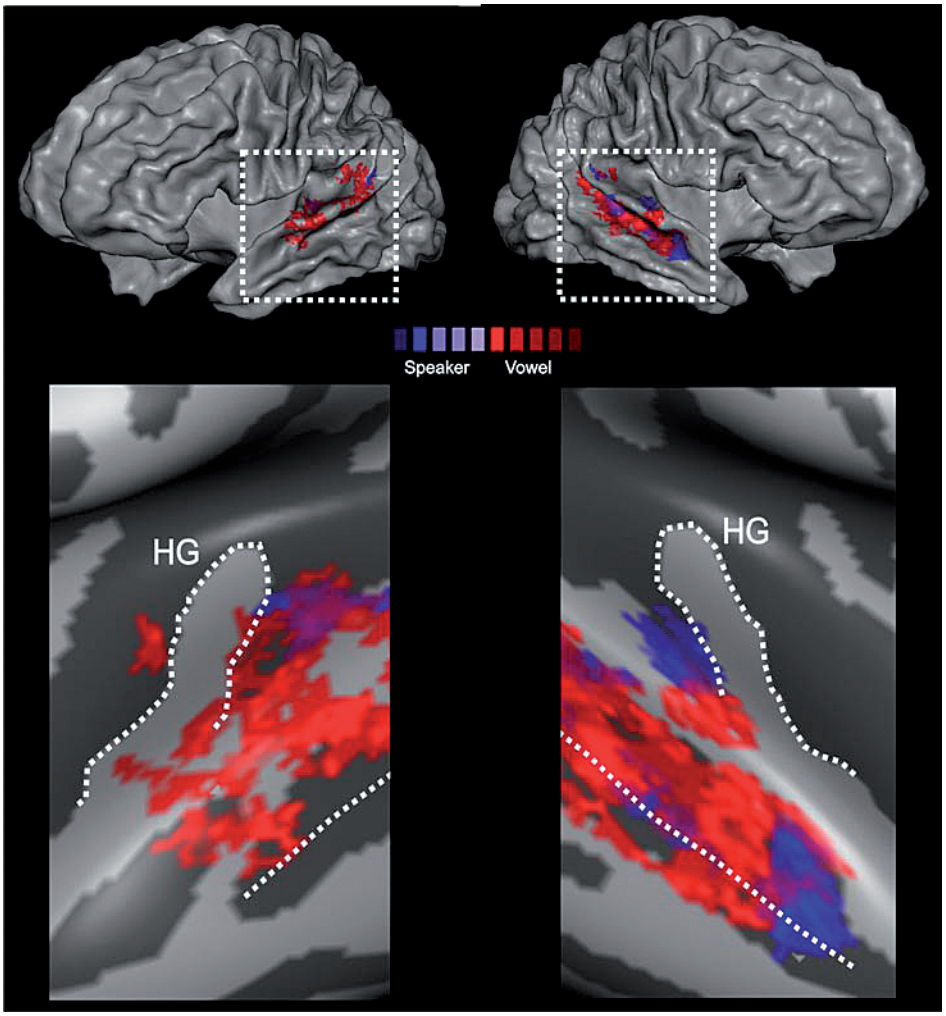

Figure 4: Superposition of the cortex-based aligned group discriminative maps for vowel (red) and speaker (blue) with a zoomed-in detail showing the relation between the informative patterns.

ing strategy, stimuli used for training and for testing differ in many acoustical dimensions. An accurate decoding of activation patterns associated with the test stimuli would thus indicate that the learned functional relation between a cortical activation patterns and a vowel (or a speaker) entails information on that vowel (or speaker) beyond the contingent mapping of its acoustic properties.

Despite the small number of training samples (10), the classification of novel stimuli was accurate in all subjects and in all possible pair wise comparisons, both in the case of vowel discrimination (/a/ vs /i/ = 0.66 (mean accuracy), $P = 1 \cdot 10^{-6}$; /a/ vs /u/ = 0.62, $P = 3 \cdot 10^{-5}$; /i/ vs /u/ = 0.60, $P = 7 \cdot 10^{-5}$; see Figure 3d) and in the case of speaker discrimination (sp1 vs sp2 = 0.62 (mean accuracy), $P = 6$

$\cdot 10^{-6}$; sp1 vs sp3 = 0.65, $P = 8 \cdot 10^{-7}$; sp2 vs sp3 = 0.63, $P = 2 \cdot 10^{-6}$; see Figure 3e). Although sparser, the corresponding discriminative maps included a subset of the locations highlighted by the previous analyses (see outlined regions in Figures 3a and 3b). These results show the capability of the brain based classifier to generalize the discrimination of vowels across speakers and of speakers across vowels. In other words, the classifier, after learning to discriminate the vowels in one speaker, was able to accurately discriminate the vowels in novel speakers. Similarly, speakers were identified also when they articulated a novel vowel.

Our findings demonstrate that estimation of the neural activation 'fingerprint' of a speech sound (vowel) is sufficient to decipher its content ('what' is being said) and the identity of the speaker ('who' is saying it). Furthermore, our results show that the auditory cortex entails representations of both 'vowel' and 'speaker' identity. The representation of 'vowel' is speaker-invariant and the representation of 'speaker' is vowel-invariant, thus supporting computational models that postulate the existence of these intermediate representations.

Based on the copious neuroimaging literature, one would have expected these representations to be sustained by a limited set of specialized and separated regions. Our findings, obtained without the constraints posed by a subtraction-based approach and with an advanced analysis strategy that allows examining localized as well as distributed response patterns, were only in partial agreement with these predictions.

The group-map of 'speaker' discrimination is in agreement with the expectations of previous studies, as it includes a set of clustered regions strongly lateralized to the right STS. These regions presumably correspond to those that have been implicated in the processing of human voices and speaker identity (Belin and Zatorre, 2003; von Kriegstein *et al.*, 2003; von Kriegstein *et al.*, 2005). In particular, the most anterior right STS cluster in our discriminative maps clearly resembles a region that has been described in a previous study that employed fMR-adaptation to investigate the (localized) neuronal representation of speaker identity (Belin and Zatorre, 2003). Possibly due to the greater sensitivity of our multivariate approach, our findings suggest that also other regions concur to the discrimination of speakers and representation of speaker identity. Conversely, our map of 'vowel' discrimination prompts for a revision of current models on processing of vowels based on the existence of localized and functionally specialized modules. Informative locations in the 'vowel' map were widely distributed and covered a large part of the activated auditory cortex in the left and right hemisphere. In the left hemisphere, discriminative patterns covered mainly the Heschl's sulcus, the planum temporale and the posterior extent of the STG, while in the right hemisphere they covered the Heschl's sulcus and the STS. Interest-

ingly, these locations include regions which have been separately reported in fMRI study that investigated processing of vowels using different sets of subtractions and control sounds (Jancke *et al.*, 2002; Liebenthal *et al.*, 2005; Obleser *et al.*, 2006; Obleser *et al.*, 2007; Scott *et al.*, 2000; van Atteveldt *et al.*, 2004). Also considering the distributed nature of the patterns obtained with the generalization analysis (see dotted lines in Fig. 3a), our findings suggest that an 'abstract' representation of a vowel emerges from the joint encoding of information occurring not only in a specialized higher level region but also in auditory areas conventionally associated with "lower-level" auditory processing. Understanding how these "lower-level" auditory areas contribute to a "higher-level" representation of a stimulus will require a detailed knowledge not only of the spatial aspects but also of the temporal dynamics of neural activations and interactions in these areas. In humans, this may be achievable by combining fMRI with high temporal resolution electroencephalography (EEG) or magnetoencephalography (MEG). In animal models, results that are compatible with our findings and interpretation have been recently reported (Mesgarani *et al.*, 2008; Nelken, 2004; Wang *et al.*, 2005).

Combining a two-factorial design with machine learning also allowed us to optimally investigate, in the same subjects and with the same stimuli, the relation between the processing of 'what'(speech) and 'who' in the auditory cortex. Note that the 'speaker' discrimination and the 'vowel' discrimination maps were obtained using identical data and an identical learning algorithm and that the only difference between the two analyses consisted in the grouping of stimuli. Also, classification accuracies were comparable for both the analyses. Observed differences in the discriminative maps thus reflect genuine differences in the informative activation patterns driving the two brain-based learning processes. While a large overlap between the two maps exists, there are also compelling differences. In particular, the presence of two adjacent speaker-informative and vowel-informative regions in the lateral portion of the right Heschl's gyrus/Heschl's sulcus (see Figure 3) suggests an early parallel processing of acoustic information relevant for speaker or vowel identification. This finding predicts the presence of an 'early' temporal marker for processing of voices similar to ones reported for processing of speech in previous EEG and MEG studies (Näätänen, 2001; Obleser *et al.*, 2004) .

In conclusion, our study demonstrates the feasibility of 'brain reading' the speech content and the speaker identity from observation of auditory cortical activation patterns of the listeners. In our experimental setting, however, speech content was as simple as a vowel, the number of sounds among which the discrimination had to be done was reduced; furthermore all sounds were presented

in isolation. Extension of our results to identify a speaker, a word or concatenation of words among many others, possibly in the context of ‘real life’ situations and complex auditory scenes provides compelling challenges for future research and experiments.

Methods

Subjects

Seven (3 females) healthy native Dutch subjects gave their written informed consent and participated in the study. None of the participants had a history of hearing loss or neurological abnormalities. Approval for the study was granted by the Ethical Committee of the Faculty of Psychology at the University of Maastricht.

Auditory stimuli

Stimuli were nine speech stimuli consisting of three natural Dutch vowels (/a/, /i/, and /u/) spoken by three native Dutch speakers (sp1: female, sp2: male, and sp3: male). We included three tokens of each vowel for each speaker, leading to a total of 27 utterances, thereby introducing some acoustic variability reminiscent of natural speech perception conditions. Stimuli were digitized at a sampling rate of 44.1 kHz, D/A converted with 16 bit resolution, band pass filtered (80 Hz to 10,5 kHz), down sampled to 22.05 kHz, and edited with PRAAT-software (Boersma, 2001). Stimulus length was equated to 230 ms (original range 172 to 338 ms), by using PSLOA (100-300 Hz as extrema for the F0 contour). Sound intensity level was numerically equated across stimuli by matching root mean square values. To avoid acoustic transients (clicks) that would be created by a sharp cut-off, stimuli were faded with 50 ms linear onset and offset ramps.

Functional MRI measurements

Brain imaging was performed with a 3 Tesla Siemens Allegra (head setup) at the Maastricht Brain Imaging Center. For each subject, two high-spatial resolution (1.5 x 1.5 x 2 mm^3) functional runs (550 volumes) were collected using a standard echo-planar-imaging sequence (repetition time [TR] =2.5 s; acquisition time [TA] =2.0 s, field of view [FOV] = 192 mm x 192 mm, matrix size = 128 x 128, echo time [TE] = 30 ms). Each volume consisted of 23 slices, covering the temporal and adjacent cortex. During the measurements, subjects listened to the stimuli that were presented binaurally and at a comfortable listening level via MR compatible headphones (Commander XG, Resonance Technology, Northridge, CA) in the 500-ms silent gap between two volume acquisitions. According to a slow event-related design, the average inter-trial-interval between two stimuli was 15 s (range 13 – 17 s). Each of the two functional runs included ten trials per stimulus condition (90 trials/run) thus resulting in twenty trials per stimulus condition. The sequence of stimuli was pseudo-randomized. Anatomical images

covering the whole brain were obtained between the functional runs using a 1 × 1 × 1 mm^3 resolution T1-weighted sequence.

fMRI Data Analysis: pre-processing and univariate statistics

Functional and anatomical images were first analysed with BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Preprocessing consisted of slice scan-time correction (using sinc interpolation), linear trend removal, temporal high-pass filtering to remove nonlinear drifts of seven or less cycles per time course, and 3-dimensional motion correction. Functional slices were co-registered to the anatomical data, and both data were normalized to Talairach space.

Conventional univariate statistical analysis of the fMRI data was based on the general linear modelling (GLM) of the time series. For each subject, a design matrix was formed using a predictor for each of the 9 stimulus condition. The predicted time courses were adjusted for the hemodynamic response delay by convolution with a canonical (double gamma) hemodynamic response function. Contrast analysis between vowels/speakers did not show any significant effect ($q = 0.05$, corrected for multiple comparison with false discovery rate).

fMRI Data Analysis: multivariate pattern recognition

Multivoxel patterns of sound-evoked BOLD responses were analysed using a novel method that combines a machine learning with an iterative, multivariate voxel selection algorithm, Recursive Feature Elimination (RFE) (De Martino *et al.*, in review). This method allows estimating maximally discriminative response patterns without a priori definition of regions of interest. In brief, starting from the entire set of measured voxels our method uses a training algorithm (least square support vector machine, ls-SVM) iteratively to eliminate irrelevant voxels and to estimate the informative spatial patterns. Correct classification of the test data increases, while features/voxels are pruned on the basis of their discrimination ability. We have recently validated and compared this method to other approaches of multivoxel pattern analysis and demonstrated its greater sensitivity by means of simulated and real data. A short description of the method is given below, together with steps and parameters specific to the analysis of present data. A more complete account of the implementation and validation of the method can be found in (De Martino *et al.*, in review). Pre-processed functional time series were first divided into “trials” (one trial per sound presentation). For each trial, a multivoxel pattern response was generated. An estimate of the response at every voxel was obtained by fitting a general linear model with one predictor coding for the trial response and one linear predictor accounting for a within-trial linear trend. The trial-response predictor was obtained by convolution of a boxcar with a

double-gamma hemodynamic response function. The corresponding regressor-coefficient (beta) was taken to represent the voxel trial response and responses from all voxels were combined to form multivoxel patterns.

In the first analysis, we labelled the stimuli and corresponding response patterns according to the relevant dimension irrespective of the other dimension, which led to the grouping of stimuli and responses in three conditions: /a/, /i/, and /u/ (*vowel learning*) and sp1, sp2 and sp3 (*speaker learning*). This resulted in a total of 60 trials per condition for vowel discrimination, and 60 trials per condition for speaker discrimination. Multivoxel pattern responses were analysed using the iterative ls-SVM-based classification algorithm. For each pair of vowels (or speakers), trials were divided into a training set (50 trials) and a test set (10 trials).

In the *generalization* analysis, the same iterative learning algorithm was trained in discriminating vowels with samples from one speaker (e.g. /a/ vs /i/ for sp1) and the accuracy of this discrimination was tested with samples from the other speakers (e.g. sp2 and sp3). Analogously, for speaker discrimination, the learning algorithm was trained in discriminating speakers with samples from one vowel (e.g. sp1 vs sp2 for /a/) and tested with samples of other vowels (e.g. /i/ and /u/).

The training set was used for estimating the maximally discriminative patterns with the iterative algorithm; the test set was only used to assess the correctness of classification of unseen trials (i.e. not used in the training). Starting from the whole-brain cortical voxels, the 2000 most active voxels per condition (as defined on the training set alone) were initially selected. Voxels were further reduced using the iterative RFE algorithm. At each iteration, RFE included two steps. First, a subset of the training data (40 trials for the first analysis, and 10 trials for the second analysis) was used to train an ls-SVM classifier. As a result of this training, a map coding for the relative contribution of each voxel to the discrimination of conditions (discriminative maps) was obtained as in (Mourao-Miranda *et al.*, 2005). Second, these discrimination weights were ranked and voxels corresponding to the smallest ranking were discarded. Voxels with the highest discriminative values were used for training in the next iteration. These two steps were repeated ten times (N_{it} = 10, on different subsets of the training data), each time with a 30% reduction in the number of voxels. The correctness of classification corresponding to the current set of voxels and the discriminative weights were assessed using the external test trials. The entire iterative procedure was repeated in cross-validation ten times (N_{splits} = 10), each time leaving out a different subset of trials per condition. For each subject and each pair of condition, the reported correctness was estimated as the average over the ten splits. Single-subject discriminative maps corresponded to the voxel-selection level that

gave the highest average correctness.

To examine the consistency of the results across subjects, group-level discriminative maps were generated after cortex-based alignment (Goebel *et al.*, 2006) of single-subject discriminative maps. In these group-level discriminative maps, a cortical location (vertex) was color-coded if it was present in the corresponding individual discriminative map of at least four of the seven subjects.

Acknowledgements

Work supported by VIDI-grant no. 452-04-337 from the Netherlands Organization for Scientific Research (NWO).

References

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8, 129-135.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14, 2105-2109.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309-312.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10, 512-528.

Boersma, 2001. Praat, a system for doing phonetics by computer. Glot International 5, 341-345.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., in review. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns.

Desai, R., Liebenthal, E., Waldron, E., Binder, J.R., 2008. Left Posterior Temporal Regions are Sensitive to Auditory Categorization. J Cogn Neurosci.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum Brain Mapp 27, 392-401.

Jancke, L., Wustenberg, T., Scheich, H., Heinze, H.J., 2002. Phonetic perception and the temporal cortex. Neuroimage 15, 733-746.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. Cereb Cortex 15, 1621-1631.

Luce, P.A., Goldinger, S.D., Auer, E.T., Jr., Vitevitch, M.S., 2000. Phonetic priming, neighborhood activation, and PARSYN. Percept Psychophys 62, 615-625.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognit Psychol 18, 1-86.

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123, 899-909.

Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns:

Support Vector Machine on functional MRI data. Neuroimage 28, 980-995.

Näätänen, R., 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 38, 1-21.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., Allik, J., Sinkkonen, J., Alho, K., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 432-434.

Nelken, I., 2004. Processing of complex stimuli and natural scences in the auditory cortex. Current Opinion in Neurobiology 14, 474-480.

Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189-234.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., Rauschecker, J.P., 2006. Vowel sound extraction in anterior superior temporal cortex. Hum Brain Mapp 27, 562-571.

Obleser, J., Lahiri, A., Eulitz, C., 2004. Magnetic brain response mirrors extraction of phonological features from spoken vowels. J Cogn Neurosci 16, 31-39.

Obleser, J., Zimmermann, J., Van Meter, J., Rauschecker, J.P., 2007. Multiple stages of auditory speech perception reflected in event-related FMRI. Cereb Cortex 17, 2251-2257.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123 Pt 12, 2400-2406.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci 26, 100-107.

Shestakova, A., Brattico, E., Soloviev, A., Klucharev, V., Huotilainen, M., 2004. Orderly cortical representation of vowel categories presented by multiple exemplars. Brain Res Cogn Brain Res 21, 342-350.

van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L., 2004. Integration of letters and speech sounds in the human brain. Neuron 43, 271-282.

von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17, 48-55.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L., 2005. Interaction of

face and voice areas during speaker recognition. J Cogn Neurosci 17, 367-376.

Wang, X., Lu, T., Snider, R.K., Liang, L., 2005. Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435, 341-346.

Predicting EEG power oscillations using fMRI

5

Abstract

The combination of different imaging modalities is emerging as a tool to study brain dynamics with both high temporal and high spatial resolution. Multimodal imaging techniques rely on the assumption of a common neuronal source for the different recorded signals. In this context, the coupling between Electroencephalographic (EEG) and functional Magnetic Resonance Imaging (fMRI) blood oxygen level dependent (BOLD) signals remains an open challenge for a useful combination of these imaging modalities. Recently, the use of simultaneous EEG-fMRI measurements has been proposed as a way to understand the relation between the two signals.

Previous attempts to the analysis of simultaneous EEG-fMRI data reported significant correlations between regional BOLD activations and modulation of EEG power, mostly in the alpha but also in other frequency bands. Beyond the correlation of the two measured brain signals, the relevant issue we address here is the ability of predicting the signal in one modality using information from the other modality.

Using multivariate machine learning-based regression, we show that is possible to predict power modulations of EEG frequency bands from simultaneously acquired fMRI data during rest (eyes open/closed) and near natural stimulation (passively watching a movie).

Based on:

De Martino, F., Valente, G., Goebel, R., Formisano, E. (2008). Predicting EEG power oscillations using fMRI. (In Preparation).

Introduction

The complementary nature of Electroencephalography (EEG; high temporal resolution and low spatial resolution) and functional Magnetic Resonance Imaging (fMRI; high spatial resolution and low temporal resolution) makes their combination appealing for investigating human brain dynamics (Dale *et al.*, 2001, Ritter *et al.* 2006, Debener *et al.*, 2006).

Blood oxygen level dependent (BOLD) responses and neural activity have been shown to be coupled (linear correlation) using simultaneous fMRI and single unit recordings in monkeys (Logothetis *et al.*, 2001), and non simultaneous intra-cortical recordings in humans (Mukamel *et al.*, 2005). Despite these results, understanding the coupling between EEG signals recorded on the surface of the scalp and the fMRI BOLD signal remains an open challenge for the useful combination of these imaging modalities.

In the last years several techniques have been proposed to combine brain signals as measured by EEG and fMRI. These methods range from the separate analysis of the data and subsequent juxtaposition of the results to truly integrated methods (Dale *et al.*, 2001). While the latter methods assume a common neuronal source for the two signals (electrophysiological; blood-oxygen-level-dependent), juxtaposition often makes the same assumption implicitly in the interpretation of the results. Integrated analysis of EEG/fMRI data can be broadly divided in two categories of methods: 1) fMRI constrained EEG analysis (equivalent current dipoles estimates and continuous current dipoles estimates); 2) EEG constrained fMRI analysis (fMRI correlates of EEG power modulations, trial-by-trial coupling).

Within this context, the development of simultaneous EEG-fMRI measurements provides several advantages over separate recordings despite the degraded EEG data quality. In particular simultaneous measurements guarantee identical sensory stimulation, perception and behaviour, and also provide a unique way to study how intrinsic brain states interact with event-related, extrinsic processing (Debener *et al.*, 2006).

Simultaneous recordings and integrated analysis techniques have been used to study the coupling between power modulations of surface EEG and fMRI BOLD signal at rest. Several studies have reported voxel-by-voxel *correlations* between BOLD changes and modulations of EEG power in different frequency bands (Goldman *et al.*, 2002, Laufs *et al.*, 2003a, Laufs *et al.*, 2003b, Moosmann *et al.*, 2003, Feige *et al.*, 2005, Mukamel *et al.*, 2005, Gonçalves *et al.*, 2006, Laufs *et al.*, 2006, de Munck *et al.*, 2007, Giraud *et al.*, 2007, Scheeringa *et al.*, 2007). This '*massively univariate*' approach does not take into account the intrin-

sic multivariate nature of the fMRI data and thus can be suboptimal in detecting relations to EEG power modulations that are weak and spatially distributed in the fMRI data.

Correlation of functionally connected networks, obtained using multivariate analysis of the fMRI data (spatial Independent Component Analysis, ICA), and the EEG power modulations in different frequency bands have been used to describe the coupling of the two measured signals (Mantini *et al.*, 2007). The data driven nature of ICA allows extracting functionally connected networks that reflect spatially independent processes. Different independent components, thus, can be coupled to the same EEG rhythm, leaving unresolved the challenging problem of finding the network in the fMRI data that explains, in a multivariate sense, most of the information in a specific EEG band.

Multiway partial least-squares (N-PLS, Martínez-Montes *et al.*, 2004) has been used to estimate the combination of voxel BOLD signals that *correlates* best with the EEG and the combination of EEG signals that *correlates* best with the BOLD signal in a unified data-driven solution that does not rely on the definition of broad EEG bands.

The use of *correlation* as the measure of the coupling between EEG and fMRI while reliable in explaining the relations between the available data sets is not optimal in generalizing the estimated relation to different data sets and does not evaluate the *predictive* power of one modality on the other.

Recently predictive models have been introduced to investigate the relation between multivariate fMRI BOLD signals and a continuous experimental variable (Pittsburgh Brain Activity Interpretation Competition (PBAIC) 2006; Formisano *et al.*, 2008, Friston *et al.*, 2008). These methods are particularly suited to the analysis of fMRI data given the typical dimensionality of the problem (number of voxels >> number of samples). Beyond simple correlation these methods allow, after a learning phase, the prediction of the experimental variable (i.e. stimulus, behaviour) exploiting the multivariate information present in the data.

In the context of simultaneous EEG-fMRI recordings the issue we address in this chapter is the ability to predict the signal in one modality using information from the other modality. In particular we predict EEG power modulations from simultaneously acquired BOLD fMRI and introduce the use of multivariate machine-learning based regression to measure the coupling between the two brain signals.

Previous studies that investigated the relation between EEG power modulations and fMRI BOLD in humans used mostly resting paradigms (Goldman *et al.*, 2002, Laufs *et al.*, 2003a, Laufs *et al.*, 2003b, Moosmann *et al.*, 2003, Laufs *et al.*, 2006, de Munck *et al.*, 2007, Giraud *et al.*, 2007, Scheeringa *et al.*, 2007)

and have reported high inter subject variability for the resulting regional correlations (Gonçalves *et al.*, 2006). We tested the ability of our method to predict EEG power modulations in a paradigm that consisted of alternating blocks of rest with eyes open and eyes closed (Feige *et al.*, 2005). Furthermore we investigated the nature of the coupling between EEG power modulations and fMRI BOLD signals during natural stimulation (passively watching a movie). Such stimulation has been shown to induce strong inter subject correlations at the level of the BOLD signal (Hasson *et al.*, 2004), and has been recently used to study the coupling between EEG (local field potentials (LFPs) and spike rates) and fMRI BOLD in non simultaneous measurements (Mukamel *et al.*, 2005).

We show that multivariate regression can predict with considerable accuracy modulation of relevant frequency bands when subjects open or close their eyes. When considering the spontaneous alpha modulations (orthogonalized with respect to the eyes open eyes closed protocol), our method replicate previously reported results for the same paradigm (Feige *et al.*, 2005) in terms of the resulting fMRI maps.

Preliminary results (one subject) on the movie sections show the ability of multivariate regression in predicting EEG power modulations elicited by complex audio-visual stimuli.

Methods

General Description of the approach

Figure 1 illustrates the main steps of our approach. The original EEG data are pre-processed in order to remove the artifacts induced by the Magnetic Resonance Imaging (MRI) environment (gradient artifact and ballistocardiogram (BCG) artifact, Allen *et al.*, 2000, Niazy *et al.*, 2005, Debener *et al.*, 2007). The pre-processed data are decomposed using temporal Independent Component Analysis (tICA; Makeig *et al.*, 1997) and interesting components are distinguished from residual artifacts based on their scalp distribution, power spectrum and event related averages with respect to the fMRI volume trigger and cardiac cycle (Debener *et al.*, 2007). EEG data are reconstructed using only the selected components and time-frequency (wavelet) decomposition is applied at the level of the single channels data or on the independent components' (ICs) time courses. Predictors are then computed averaging the power modulations in different frequency windows (delta, theta, alpha, beta and gamma) and convolving with a canonical hemodynamic response function (HRF; Friston *et al.*, 1998).

Data sets (i.e. the pre-processed (see below) fMRI time series and EEG based predictors), are then divided into training and testing. The training data set is used to learn multivariately (using Relevance Vector Machine and Ridge Regression) the coupling between the fMRI and the EEG modulations in the selected frequency band. A map depicting the contribution of each voxel to the learned coupling is obtained. The test data are used to assess the validity of the learned coupling.

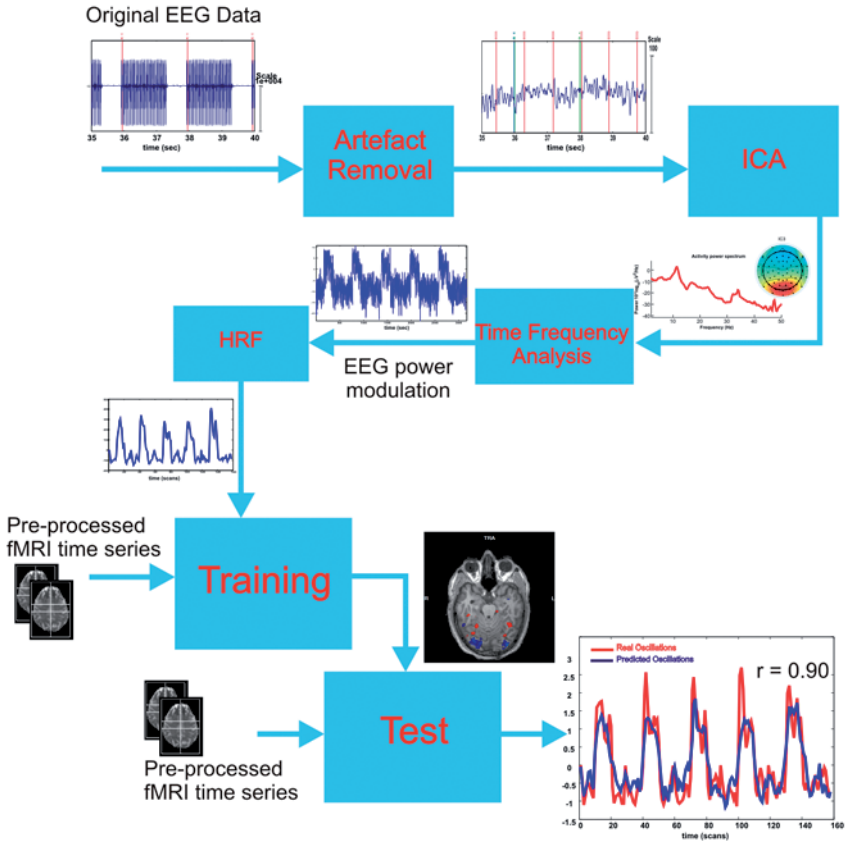

Figure 1: General description of the proposed approach to the predicion of EEG power oscillations frm fMRI data. The original EEG data are first pre-processed in order to remove the MRI related artifacts. Independent Component Analysis is used to select components of interest. Time frequency analysis is performed on the selcted IC time courses and a predictor for the analysis of the fMRI data is obtained after convolving with a canonical hemodynamic resonse function. Multivariate regression is then performed training on a subset of the available data and evaluating (test) the performances on the remining data.

Multivariate Regression of fMRI time series

In the context of simultaneous EEG-fMRI recordings, the fMRI time series, represented by the *N* x *V* matrix **X** (*N* being the number of volumes and *V* the number of voxels), and the EEG power modulations in a specific band, represent-

ed by the N dimensional vector $\mathbf{o}$, are employed as a data set D. Such dataset can therefore be seen as a collection of N pairs $(\mathbf{x}_i, o_i)$, denoting with $\mathbf{x}_i$ a sample vector of dimension V (one volume of the fMRI time series) and with o_i the corresponding one-dimensional label (the EEG power modulations in a specific band as extracted from the simultaneous recordings).

In what follows we describe the estimation of linear generative models used for multivariate regression of fMRI time series.

Linear Model Estimation

Linear models have been extensively employed in machine learning for fMRI data analysis. There are many reasons for this choice. The high number of voxels in fMRI datasets, in fact, poses some challenges to the model estimation. If no feature reduction is performed the number of voxels V (i.e. dimension of the feature space) is usually considerably higher than the amount of samples N, and therefore the effects of the curse of dimensionality are not negligible. Furthermore, when using linear models, is straightforward to obtain a map that helps understanding the relevance of different brain regions to the prediction.

A standard linear model has the following form:

$$o = \mathrm{y}(\mathbf{x}, \mathbf{w}) + \varepsilon \qquad (1)$$

where $\mathrm{y}(\mathbf{x}, \mathbf{w})$ is the deterministic input-output mapping part and ε accounts for the noise in the measurements. The deterministic mapping can be modelled as (Bishop, 2006):

$$\mathrm{y}(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_V x_V = \mathbf{w}^T \tilde{\mathbf{x}} \qquad (2)$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_V)^T$ denotes the training dataset (defined over a V-dimensional space), $\tilde{\mathbf{x}} = (1, \mathbf{x}^T)^T$ and the V+1-dimensional vector $\mathbf{w}$ indicates the weights of the linear model (with w_0 indicating the bias term). This model, simply known as *linear regression* is widely employed in fMRI data analysis.

However, for the purpose of generalization, we will consider now a mapping

$$\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^T \qquad (3)$$

with $\phi: \Re^V \rightarrow \Re^M$, mapping the V-dimensional space of $\mathbf{x}$ into an M-dimensional one. ϕ can be for instance a linear polynomial, or radial basis function. Eq. (2) then becomes:

$$\mathrm{y}(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \phi(\tilde{\mathbf{x}}) = \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) \qquad (4)$$

where this time $\mathbf{w}$ is an M-dimensional vector of parameters.

The aim of the estimation procedure is to find the "best" model parameters $\mathbf{w}$, evaluating the performances in terms of an unknown dataset (generalization). One criterion could be to maximize the fit of the model to the training data. Anyway, this procedure may be dangerous, as complex models may fit also the noise

term.

One common error function in the case of regression is the sum of squares:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{o_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 \quad (5)$$

The minimization of this function (setting its gradient to zero) leads to the following estimate of the model parameters:

$$\tilde{\mathbf{w}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T o \quad (6)$$

where $\boldsymbol{\Phi} = (\phi(\mathbf{x}_1),..., \phi(\mathbf{x}_N))^T$. It can be shown that the least-squares solution corresponds to the projection of the target o onto the subspace generated by the columns of $\boldsymbol{\Phi}$ (Bishop, 2006). As discussed previously, a perfect fit on the training dataset may not be optimal for generalization purposes, therefore some *regularization* coefficients are introduced, to control for the smoothness of the estimate. The new error function will then be:

$$E(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \quad (7)$$

where E_D is the same as in Eq. (5) and E_W is the regularization term. A simple form of regularizing term is the following (Bishop, 2006):

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad (8)$$

that leads to the solution:

$$\tilde{\mathbf{w}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I})^{-1}\boldsymbol{\Phi}^T o \quad (9)$$

that is sometimes called *ridge regression* solution. Regularization is particularly effective in training on small datasets (reducing the model complexity and subsequently the risk of overfitting), but one has to employ a suitable value for the weighting coefficient λ. One way to set this parameter is to perform cross-validation choosing the parameter that gives the highest generalization on the validation set(s).

Relevance Vector Machine Regression

Similarly to Support Vector Machines (SVM; Vapnik, 1995), this method is based on the linear combination of kernel functions, with one kernel associated with each data point (in the training dataset). Compared to SVM, RVM provides in many applications a much sparser model, typically an order of magnitude more compact, with little or no reduction of generalization error. Furthermore, no parameter has to be estimated in cross-validation (like C and ε in SVM; Tipping, 2001).

The kernel can be defined starting from a nonlinear feature space mapping $\phi(\mathbf{x})$, as in Eq. (3):

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}') \quad (10)$$

The model, with N+1 parameters, can be written as:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{n=1}^{N} w_n k(\mathbf{x}, \mathbf{x}_n) + b \quad (11)$$

with b being the bias term and $k(\mathbf{x}, \mathbf{x}_n)$ the kernel function centered on $\mathbf{x}_n$.

Considering Eq. (1), and assuming that the noise term ε follows an independent, identically distributed Gaussian distribution with zero mean and precision (inverse of the variance) equal to β, that is $p(\varepsilon|\beta) = N(0, \beta^{-1})$, it follows that:

$$p(\mathbf{o}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} N(\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \tag{12}$$

where with **o** we denote the vector of all the N targets (e.g. the EEG power modulations in a specific band as extracted from simultaneous recordings).

Using Bayes' rule it is possible to express the probability of the model parameters:

$$p(\mathbf{w}|\mathbf{o}, \mathbf{x}, \beta) = \frac{p(\mathbf{o}|\mathbf{w}, \mathbf{x}, \beta)\, p(\mathbf{w})}{p(\mathbf{o}|\mathbf{x}, \beta)} \tag{13}$$

that is:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \tag{14}$$

where the prior term contains all the information one has on the model parameters. For a review of Bayesian methods refer to (Duda *et al.*, 2001, Bishop, 2006).

In RVM the prior on the model weights **w** is:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{N+1} N(0, \alpha_i^{-1}) \tag{15}$$

with an *hyperparameter* α_i for each model weight. It can be shown that the posterior distribution of the weights is again Gaussian (Tipping, 2001), with mean and covariances given by:

$$\mathbf{m} = \beta \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{o} \tag{16}$$

$$\boldsymbol{\Sigma} = (\mathbf{A} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \tag{17}$$

with $\boldsymbol{\Phi}$ being the kernel matrix as in eq. (6) and $\mathbf{A} = \text{diag}(\alpha_i)$. The estimation of incorporates the *Automatic Relevance Determination* (ARD) (MacKay, 1994, Neal, 1996). In fact, during the training phase, many hyperparameters α_i will grow to infinity, so that the corresponding model weight w_i will have a posterior distribution concentrated around zero. In other words, only the model weights (and therefore the functions associated with these parameters) that are "relevant" given the training data will remain, pruning out the unnecessary ones and leading to a sparse model. *Relevance vectors* can be seen as similar to the *support vectors* in the SVM formulation.

The values of α_i and β are determined using type-II Maximum Likelihood (known also as *evidence approximation*) (Bishop, 2006, Tipping, 2001). Once these parameters have been estimated, the prediction over a new data point can

be done averaging across *all* the possible models weighted by their probabilities. In other words, considering a new data point $\mathbf{x}^*$ then the predicted value $\mathbf{t}^*$ will be distributed according to the following:

$$p(o^*|\mathbf{x}^*, \mathbf{x}, \mathbf{o}) = \int p(o^*|\mathbf{x}^*, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{o}) d\mathbf{w} \quad (18)$$

and considering a predictive distribution that is still Gaussian, with mean and variance given by (Bishop, 2006, Tipping, 2001):

$$\mathbf{m}(\mathbf{x}^*) = \mathbf{m}^T \phi(\mathbf{x}^*) \quad (19)$$

$$\sigma^2(\mathbf{x}^*) = (\beta)^{-1} + \phi(\mathbf{x}^*)^T \Sigma \phi(\mathbf{x}^*) \quad (20)$$

Without loss of generality, we refer to the training of the model on the first functional run. Suppose that both the time courses of the voxels in the fMRI data and the simultaneously recorded EEG power modulations have zero mean (note that the evaluation metric is based on Pearson correlation). Following the formulation proposed in Eq. (11), and considering a linear kernel, we have the following model:

$$y(\mathbf{X}_1, \mathbf{w}) = \mathbf{X}_1 \mathbf{X}_1^T \mathbf{w} = \mathbf{K}\mathbf{w} \quad (21)$$

with $\mathbf{w}$ ($n_1 \times 1$) being the model weights vector and $\mathbf{K} = \mathbf{X}_1 \mathbf{X}_1^T$ ($n_1 \times n_1$) the linear kernel constructed from the starting training dataset $\mathbf{X}_1$ ($n_1 \times v$).

The RVM training aims at finding an estimate of the posterior distribution of the weights $\mathbf{w}$. This posterior distribution can be then used to perform predictions on a new dataset (second functional run) by means of Eq. (18). Denoting with $\tilde{\mathbf{w}}$ the estimated posterior mean, then the mean of the predictive distribution (Eq. (19)) is then:

$$\tilde{\mathbf{o}}_2 = \mathbf{X}_2 \mathbf{X}_1^T \tilde{\mathbf{w}} \quad (22)$$

where $\tilde{\mathbf{o}}_2$ ($n_2 \times 1$) is the estimate of the ratings on the second dataset $\mathbf{X}_2$ ($n_2 \times v$). It is possible, considering Eq. (22), to express the prediction in terms of maps:

$$\tilde{\mathbf{o}}_2 = \mathbf{X}_2 \tilde{\mathbf{M}} \quad (23)$$

with

$$\tilde{\mathbf{M}} = \mathbf{X}_1^T \tilde{\mathbf{w}} \quad (24)$$

where $\tilde{\mathbf{M}}$ ($v \times 1$) can be interpreted as a map of relative contribution of the different voxels to the final prediction.

Data

Simultaneous EEG-fMRI data were collected from four subjects, three runs per subject. Each run consisted of four five minutes movie segments alternated with one minute rest periods in which subjects were asked to close and open their eyes every twenty seconds.

EEG data

EEG data were recorded in the MRI environment using a 64-channel high-input impedance amplifier system specifically designed (Brainproducts, Munich, Germany). The setup consisted of two 32-channel MR plus amplifiers powered by a rechargeable power unit. The amplifiers were placed directly behind the scanner bore inside the MR room, which allowed the use of short wires with a total length of about 1.2 m from recording electrodes to amplifier. Sintered Ag/AgCl ring electrodes with built-in 5 kΩ resistors were used. Data were recorded from 62 equidistant scalp sites mounted in a cap system (Easycap, Falk Minow Services, Herrsching, Germany). Additional plastic electrode holders were tied into the cap at occipital scalp sites which substantially improved subject comfort. Continuous data were also recorded from one electrode placed below the left eye to monitor eye blinks and another electrode placed at the lower back for electrocardiogram (ECG) recording. All 64-channel data were referenced to the vertex. The data were recorded with a pass-band of 0.016–250 Hz and digitized with 5000 samples/s at 16-bit resolution, resulting in a dynamic range of 16.38 mV. The amplified signal was transmitted via fiber-optic cables to a recording PC placed outside the MR room. Electrode impedances were maintained below 20 kΩ before recordings.

The EEG data processing was carried out using EEGLAB (Delorme *et al.*, 2004). fMRI gradient artifacts were removed using the EEGLAB plug-in FMRIB 1.21 (Niazy *et al.*, 2005), as developed by the Centre for the Functional MRI of the Brain (Oxford, UK). The algorithm for gradient artifact removal combines template subtraction methods (sliding window of 60 artifacts) to optimal basis set (OBS) and adaptive noise cancellation (ANC) for the removal of residual artifacts after the subtraction of the template. Prior to the ballistocardiogram (BCG) artifact correction the EEG data were resampled at 250 Hz. In order to remove the BCG artifact the continuous data are epoched relative to the heartbeat events, detected automatically using the ECG electrode, and aligned in a matrix to calculate the first three principal components, which are then taken to form the OBS. The OBS are then least-squares-fitted and subtracted from each segment. After removal of the imaging related artifacts, data were filtered between 0.5 and 50 Hz, ECG and EOG channels were discarded. After artifact correction each run was divided into two data sets, the first containing the movie segments and the second containing the rest periods (eyes open and eyes closed segments). Temporal independent component analysis was used to extract 62 components separately for the movie and rest data sets after concatenating the data from the three runs. The independent components ICs were reviewed in order to discard components reflecting residual BCG or gradient artifacts (Debener *et al.*, 2007). Interesting

ICs, selected based on their scalp distribution and power spectrum, were used to reconstruct the EEG data of the three different runs.

Time frequency decomposition (Morlet wavelets) was applied to the reconstructed channel level data or the IC time courses in order to extract the power modulations in the frequency interval [1 – 50 Hz]. IC power modulations were obtained averaging the time frequency decomposed data in different frequency windows, selected around peaks of the power spectral density of each component. In addition to these "narrow band" modulations, a broad band modulation was considered for all ICs averaging the time frequency data in the window [1 – 20 Hz].

A predictor for the analysis of the simultaneously acquired fMRI data was obtained for all ICs by convolving the different power modulations with a standard hemodynamic response function (HRF; Friston *et al.*, 1998) and re-sampling the data to the fMRI sampling rate (0.5 Hz, see below).

In addition to the predictors accounting for power modulation in specific frequency bands a predictor for the analysis of the fMRI data was obtained by convolving the IC time courses root-mean-square (RMS) with a standard HRF and re-sampling to the fMRI sampling rate.

fMRI data

Functional magnetic resonance time series were acquired in a 3T system (Siemens Allegra). Functional runs consisted of 22 axial slices obtained with a T2-weighted gradient echo, EPI sequence (TR 2 s; TA 1.4 s; FOV 224 x 224; matrix size 64 x 64, voxel size = 3.5 x 3.5 x 4 mm^3). Anatomical images were obtained using a high resolution (1 x 1 x1 mm^3), T1-weighted sequence.

The fMRI data sets were subjected to a series of pre-processing operations. (1) Slice-scan-time correction was performed by resampling the time courses with linear interpolation such that all voxels in a given volume represent the signal at the same point in time. (2) Head movements were detected and automatically corrected by minimizing the sum of squares of the voxel-wise intensity differences between each volume and the first volume of the run. Each volume was then resampled in three-dimensional space according to the optimal parameters using trilinear interpolation. (3) Temporal filtering was performed in order to remove linear trends from the voxels' time series. (4) After co-registration to the anatomical images collected in the same session the functional volumes were projected into Talairach space.

After pre processing of the time series each run was divided into two data sets, the first containing the movie segments and the second containing the rest periods (eyes open and eyes closed segments). These data sets were separately

used for the multivariate regression analysis using both Ridge Regression (RR) and Relevance Vector Machines (RVM). Training was performed using the data of run 1 and 3 and predictions were evaluated on the data of run 2. When using RR the number of features/voxels to be used was cross-validated during the training phase using correlation based ranking of the voxels.

Multivariate regression was compared to a conventional univariate regression analysis (General Linear Model, GLM) and to a massively multivariate analysis performed using Independent Component Analysis (ICA). GLM analysis was performed concatenating run 1 and 3 and estimating voxels' beta weights for each EEG related predictor. The estimated beta weights were then used to predict the EEG power modulations in run 2. In order to predict power EEG power modulations using fMRI ICA analysis, the independent components (ICs) with the highest correlation to each EEG related predictor were identified in run 1 and 3. Using the similarity of the ICs spatial maps (Esposito *et al.*, 2005) a candidate IC was selected to predict the EEG power modulations in run 2, and the correlation between the selected IC time course and the EEG power modulations was used as a measure of prediction.

For visualization of resulting predictive maps, folded cortex was extracted from anatomical MRI data and used to calculate flat maps.

Results and Discussion

Figure 2 shows single subject (FDM) results obtained for the prediction of the EEG power modulations of an occipital independent component (figure 2 top left) extracted during the eyes open/closed paradigm. All methods were trained on the modulation of the power in the frequency band [8 – 14 Hz] (alpha band; figure 2 top right) of run 1 and 3 and predictions are reported for run 2.

All methods achieved considerable prediction results (reported as the correlation between the real and predicted EEG power modulations; figure 2), with higher accuracies obtained with multivariate regression based on Relevance Vector Machines (RVM).

Figure 2 also shows the maps, projected on the flattened cortex of the subject, obtained as a result of the training for all methods. Despite the different estimation procedure, all methods highlighted the same occipital regions as the most relevant for prediction.

Conventional statistical analysis of combined EEG-fMRI measurements is based on the voxel-by-voxel hypothesis testing (General Linear Model, GLM).

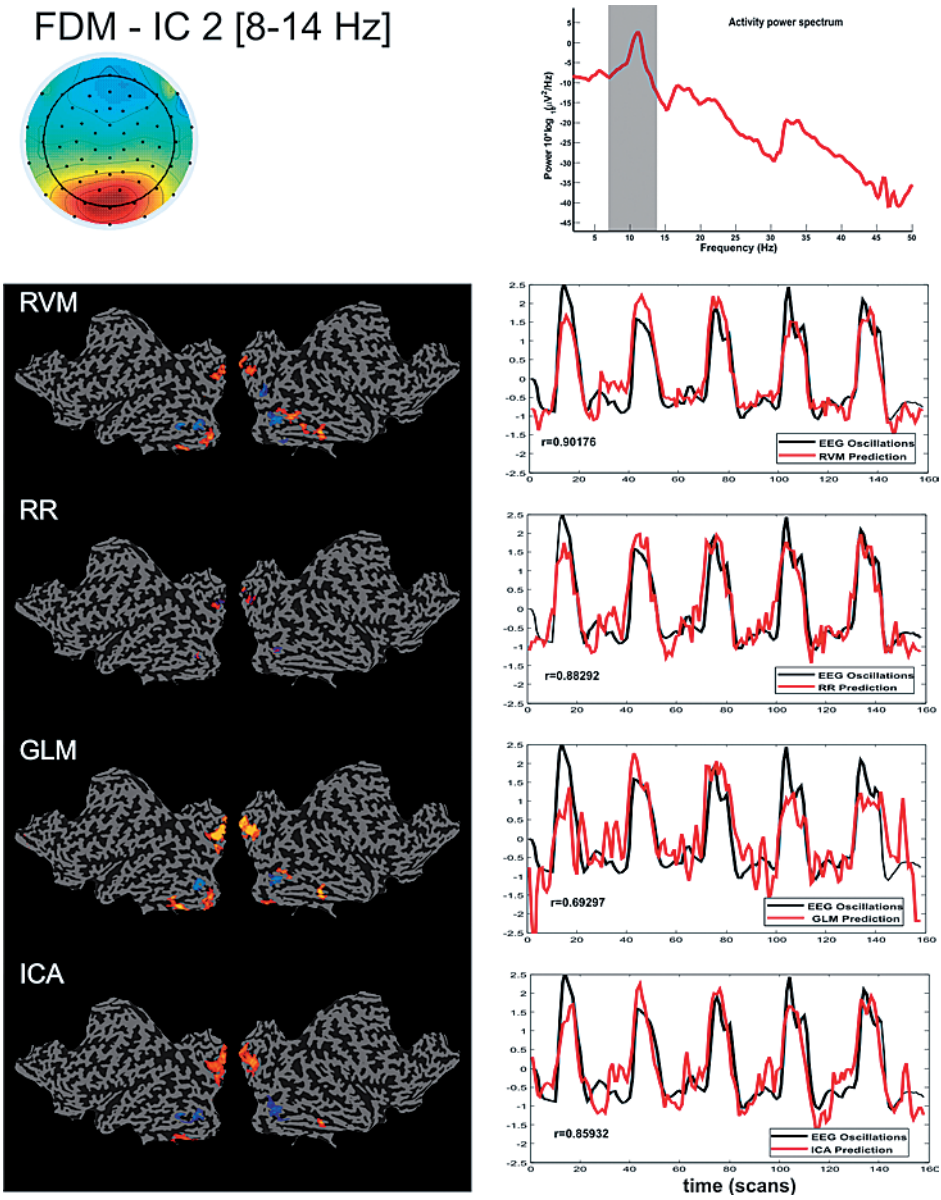

Figure 2: Results obtained for the prediction of the eyes open eyes closed induced alpha modulation in subject 1 (FDM). The selected IC topography is shown on the top right. On the top left the power spectrum of the selected IC time course is presented. Predicted (red) and true (back) EEG power oscillations in the alpha band in the second run are presented on the bottom left for all used method (RVM, RR, GLM, ICA). The maps obtained as a result of the training of each model are shown on the bottom right.

The maps obtained with such procedure depict the voxels in the fMRI data that "best fit" the modelled EEG power modulations and are estimated trying to minimize the fitting error. Such explanatory procedure may be sub-optimal when EEG

power modulations are encoded in a multivariate and distributed fashion in the fMRI BOLD signal. Furthermore the estimation procedure at the basis of the GLM while optimal to explain the available data (training set), might fail when the purpose is to generalize to unseen data (test set).

Multivariate analysis techniques such as ICA have been previously used to analyse simultaneous EEG-fMRI recordings. Independent Components (ICs) maps depict functionally connected brain regions and are estimated in a data-driven way without making use of the available experimental variables. Prediction ability of ICA is evaluated post-hoc as the correlation of a selected IC time-course with the EEG power modulations and thus can be considered as "incidental" as it is not at the basis of the estimation procedure.

Multivariate regression is an appealing tool for investigating the relevant question of the coupling between the fMRI time series and simultaneously acquired EEG modulations. Given the massive multivariate nature of the problem the use of "regularized" models such as Ridge Regression (RR) and Relevance Vector Machines (RVM) is necessary in order to avoid "overfitting" of the data. These methods aim at optimizing prediction abilities to unseen data sets thus reducing the fit of the training set. RR and RVM maps thus show the brain network that (multivariately) is most relevant in generalizing the learned EEG-fMRI coupling.

The highlighted positive and negative coupling between EEG protocol induced alpha modulations and BOLD fMRI in the occipital cortex was to be expected and has been previously reported (Feige *et al.*, 2005). No sub-cortical regions were identified as relevant for the coupling between the fMRI BOLD signal and the protocol induced alpha modulations.

Table 1

	RR	RVM	GLM	ICA
FDM; IC 7 (8 - 14 Hz)	0.85 (0.20)	0.90 (0.32)	0.63 (0.008)	0.83 (0.06)
FDM; IC 2 (8 - 14 Hz)	0.88 (0.21)	0.90 (0.26)	0.69 (0.24)	0.85 (0.05)
FM; IC 6 (7 - 14 Hz)	0.79 (0.35)	0.81 (0.35)	0.39 (0.33)	0.69 (0.21)
MR; IC 8 (8 - 14 Hz)	0.68 (0.17)	0.74 (0.23)	0.36 (0.002)	0.22 (0.23)
TP; IC 5 (7 - 14 Hz)	0.75 (0.39)	0.81 (0.23)	0.62 (0.24)	0.46 (0.26)

Accuracies obtained with Ridge Regression (RR), Relevance Vector Machines (RVM), General Linear Model (GLM) and Independent Component Analysis (ICA) in the prediction of the eyes open eyes closed power modulation of selected EEG Independent Components (IC) in the [7-14 Hz] band. In brackets the prediction of the power modulation after orthogonalization with respect to the eyes open/close protocol.

To investigate the coupling between the fMRI BOLD signal and spontaneous

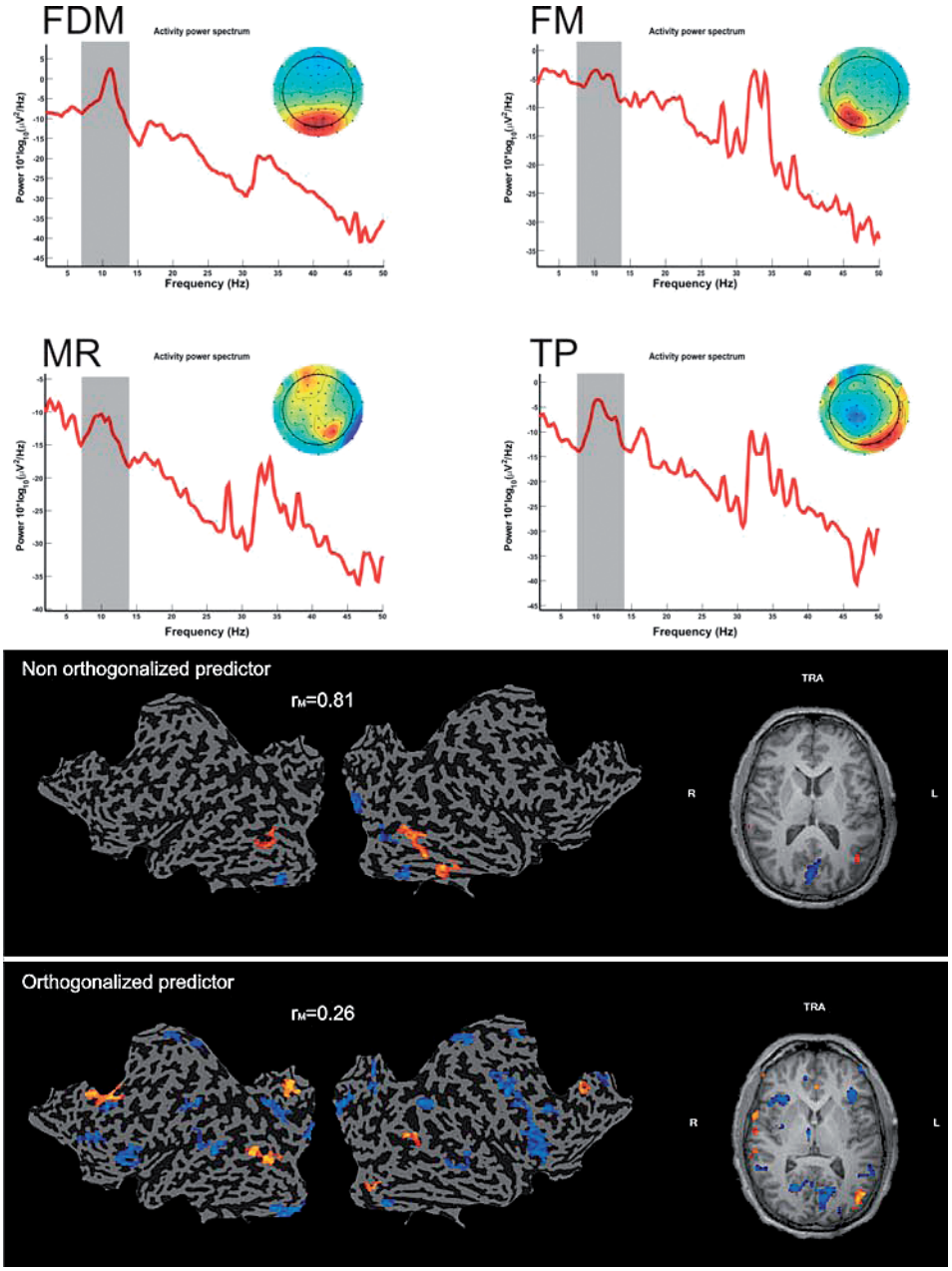

Figure 3: Group results obtained during the eyes open/closed paradigm. On the top, topographies of all subjects' ICs selected as characterizing the alpha modulation are prsented together with the power spectrum of their time courses. On the bottom, the group maps obtained after training the RVM on run 1 and 3 are shown for both the induced oscillations and the oscilaltions after orthogonalization with respect to the protocol. Mean (across subjects) predictive accuracies are reported for both cases.

EEG alpha modulations we investigated the ability of the different methods in predicting the EEG alpha predictor after orthogonalization with the eyes open/closed protocol (Feige *et al.*, 2005).

Table 1 reports, for all subjects, prediction results obtained using the same methods (RR, RVM, GLM and ICA) on the EEG independent components (ICs) that mostly reflected the alpha power modulations induced during the eyes open/closed periods. We also report between brackets the prediction results for the orthogonalized alpha predictor. Relevance Vector Machines shows the highest prediction accuracies especially for the non-orthogonalized alpha predictors.

Figure 3 shows the group results for the eyes open/close periods. The selected EEG-ICs for all subjects together with their power spectrum are presented on the top. On the bottom group maps are reported on the flattened cortex of one subject together with one transversal slice for both the protocol induced alpha modulations (non-orthogonalized predictor) and the spontaneous fluctuation (orthogonalized predictor). Group maps were obtained averaging the individually trained RVM maps in Talairach space after binarization. The colour code represents the probability of finding each voxel in all subjects together with the sign of the coupling between EEG alpha modulations and BOLD signal. Group maps were thresholded in order to show voxels present in at least half of the subjects.

Consistently with the single subject result the group map shows regions in the occipital cortex contributing positively and negatively to the prediction of the protocol induced alpha modulations. When analysing the coupling between fMRI BOLD and the spontaneous fluctuations of alpha (orthogonalized predictor), a widespread negative contribution to the prediction was found in several cortical areas (occipital cortex, parieto-frontal network, insula) as already reported in previous combined EEG-fMRI studies (Goldman *et al.*, 2002, Laufs *et al.*, 2003a, Laufs *et al.*, 2003b, Moosmann *et al.*, 2003, Gonçalves *et al.*, 2006, Laufs *et al.*, 2006, de Munck *et al.*, 2007). Interestingly in line with the findings of Feige *et al.* (2005), a negative contribution to the prediction of the spontaneous alpha fluctuations was also highlighted in the thalamus (orthogonalized predictor, transversal slice). Positive contribution to the prediction of the alpha modulations is visible in the pre-cuneus parietal cortex and anterior cingulate cortex. These regions have been consistently reported to de-activate during task execution (activate during rest) and form the “default mode” network (Greicius *et al.*, 2003), which has been previously reported to positively correlate with the alpha rhythm (Mantini *et al.*, 2007).

Following the analysis of the eyes open/closed part of the experiment we trained (run1 and 3) a Relevance Vector Machine to predict (run 2) the EEG power modulations elicited during continuous stimulation (free watching a movie).

Figure 4 – 6 report preliminary results obtained for the prediction of power modulations of different independent components extracted from the EEG data of one subject (FDM). For each component we report the scalp distribution, the power spectrum of the IC time-course together with the different frequency bands used to generate predictors for the combined EEG-fMRI analysis. We also show the maps obtained as a result of the RVM training on run 1 and 3 depicting the most relevant voxels used for the prediction and the predicted and real EEG power modulations of run 2.

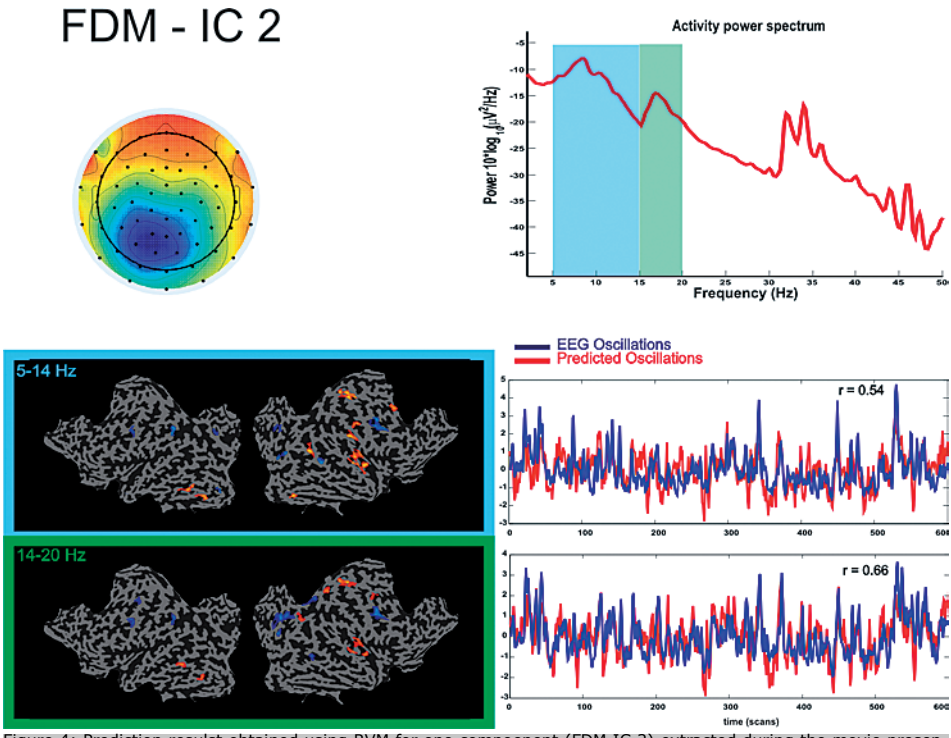

Figure 4: Prediction resulst obtained using RVM for one compoenent (FDM IC 2) extracted during the movie presentation. Topography and power spectrum of the time course of the component are shown on the top. On the bottom predicted (red) and real (blue) oscillations are shown on the right together with the map resulting from the training (run1 and 3). Resulst are reported for two selected frequency bands.

Figure 4 shows an independent component (IC) with clear parieto-occipital topography possibly reflecting subjective modulations of a posterior attention network. When trying to predict the EEG power modulations in the frequency bands [5 – 14 Hz] and [15 – 20 Hz] using RVM based regression we obtained considerable prediction accuracies (0.54 and 0.66 respectively on 600 points). The fMRI maps associated with these predictions show a negative contribution of the fronto-parietal network especially for the modulations in the [15 – 20 Hz] band. Voxels contributing positively to the prediction are visible in the ventral

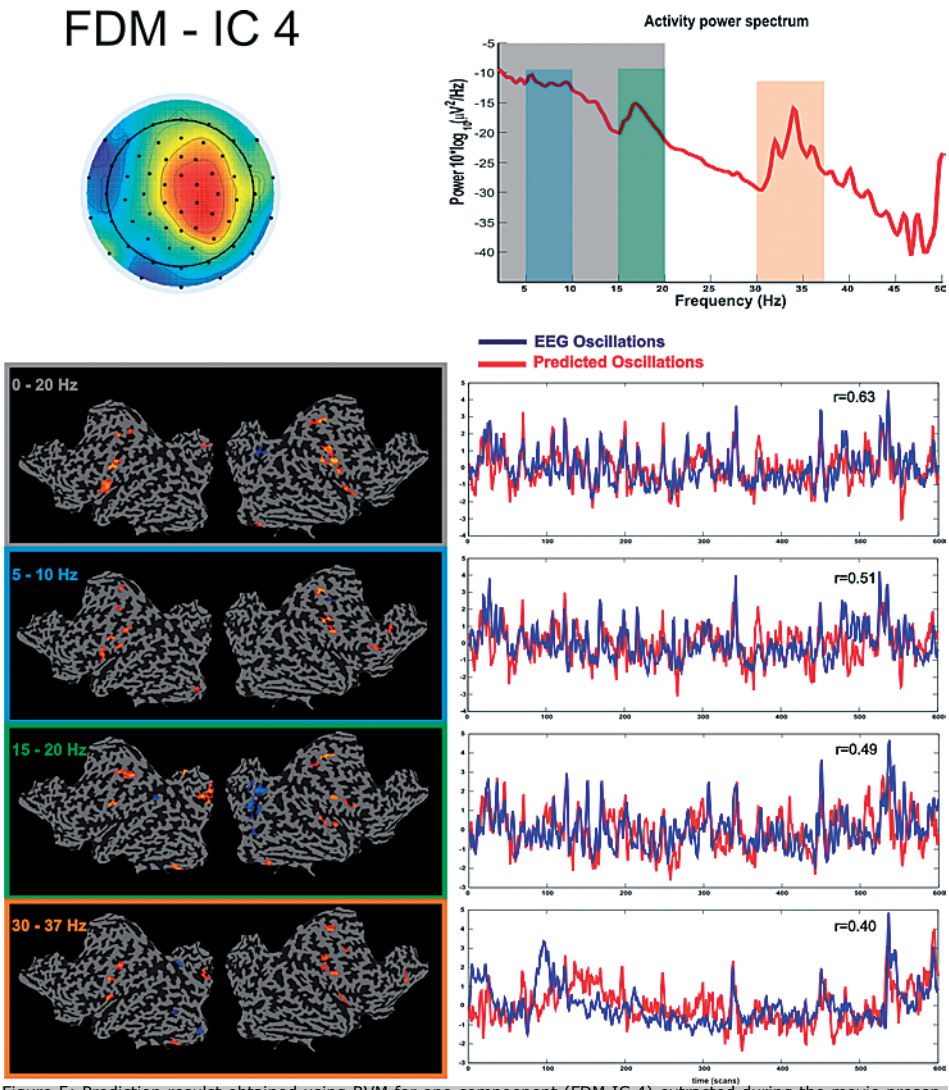

Figure 5: Prediction resulst obtained using RVM for one compoenent (FDM IC 4) extracted during the movie presentation. Topography and power spectrum of the time course of the component are shown on the top. On the bottom predicted (red) and real (blue) oscillations are shown on the right together with the map resulting from the training (run1 and 3). Resulst are reported for four selected frequency bands.

visual cortex and anterior insula.

Figure 5 shows the prediction results for a component with a topography indicating involvement of the motor cortex of the subject. The power spectrum of the IC time course shows a clear peak around 17 Hz. RVM based prediction achieved considerable accuracies for all selected spectral windows ([0 – 20 Hz], r=0.63; [5 – 10 Hz], r=0.51; [15 – 20 Hz], r=0.49; [30 – 37 Hz], r=0.40). The RVM maps obtained as a result of the training show a positive contribution from the

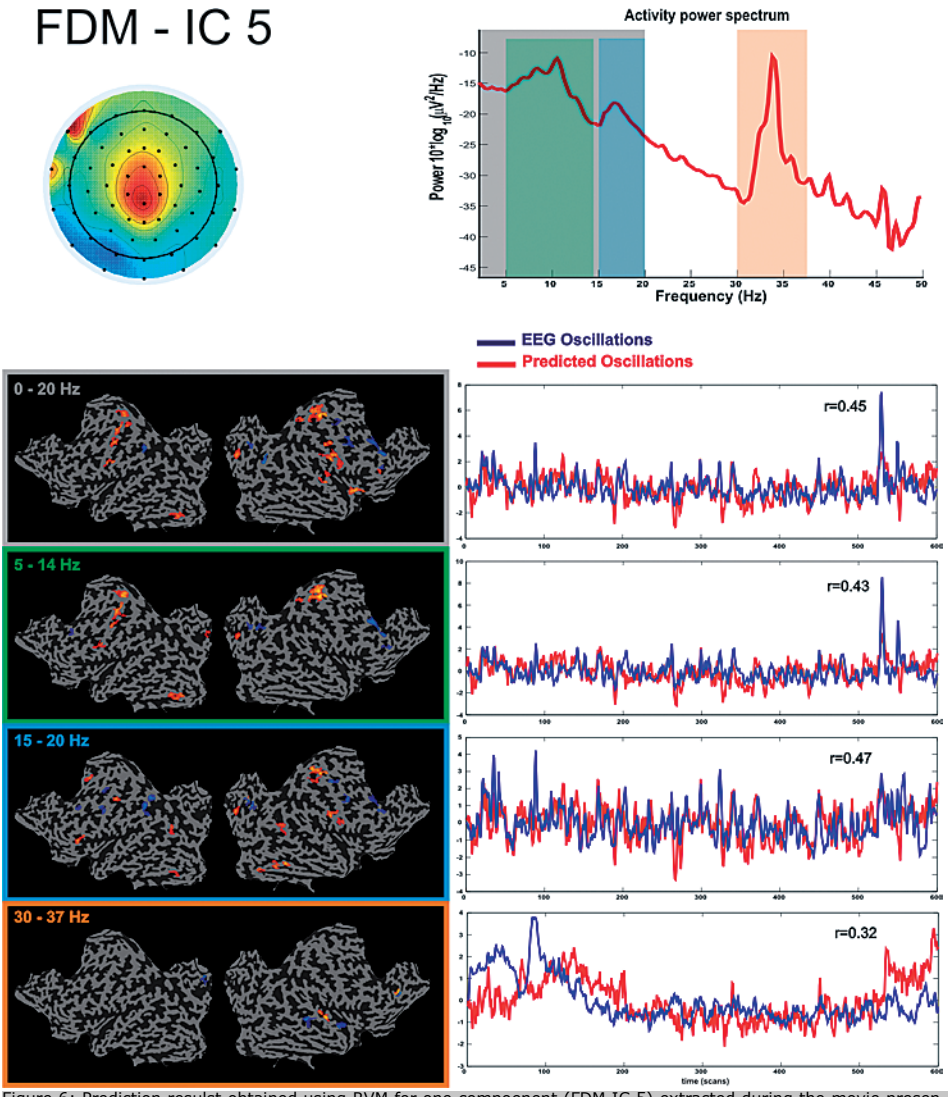

Figure 6: Prediction resulst obtained using RVM for one compoenent (FDM IC 5) extracted during the movie presentation. Topography and power spectrum of the time course of the component are shown on the top. On the bottom predicted (red) and real (blue) oscillations are shown on the right together with the map resulting from the training (run1 and 3). Resulst are reported for four selected frequency bands.

bilateral motor cortex of the subject for all selected frequency bands.

The results obtained for the prediction in the [15 – 20 Hz] band are in line with previously reported source localization experiments of the mu-rhythm (Salmelin *et al.*, 1994, Caetano *et al.*, 2007). The mu-rhythm, besides being associated with voluntary movements, has been identified as one spontaneous rhythm present in subjects during rest (Salmelin *et. al.*, 1994), and has been also associated with passive observation of actions (Caetano *et al.*, 2007). In order to disentan-

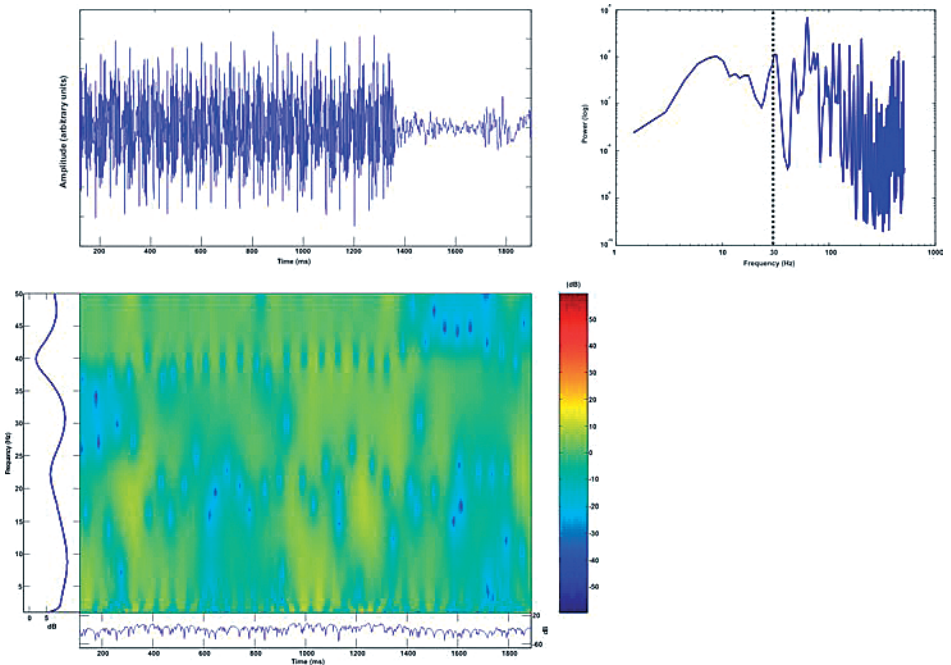

Figure 7: Analysis of the sound produced by the scanner during acquisition of one volume (top left). The power spectrum is presented on the top right. On the bottom left time freqency analysis (limited between 0 - 50 Hz) of the scanner noise.

gle the contribution to identified mu-rhythm between voluntary movements and modulations elicited by the observation of the movie, objective measurements (electromyography) of the subject's movements during the experimental protocol are needed.

Figure 6 shows the results obtained for the prediction of frequency modulations extracted from the time course of a component with topography reflecting sources in the auditory cortex (bilateral). When learning the association between the modulations in the [30 – 37 Hz] band and the fMRI data the RVM identified as relevant several regions in the primary (positive contribution) and secondary (negative contribution) right auditory cortex of the subject. Interestingly when analysing the characteristic sound produced by the scanner during the acquisition of one volume (TR) (figure 7) a clear peak in the same frequency band ([30 – 37 Hz]) is visible.

Conclusions

We have shown that multivariate regression can be used to predict power modulation of EEG frequency bands both in the case of simple induced alpha modulations (eyes open/ closed paradigm) and in the case of complex audio-visual stimuli (movie).

When comparing the accuracies obtained when predicting the alpha modulations induced when the subjects close their eyes, univariate models (GLM) and massive multivariate models (ICA) result in similar but weaker performances with respect to predictive models (RR, RVM). In particular on our data Relevance Vector Machines based prediction shows to be the most accurate. These results can be explained by the predictive (compared to explanatory or data-driven) and massive multivariate (compared to univariate) nature of RVM based regression.

Further analysis is needed to evaluate the data obtained on the movie sections. In particular the use of inter subject correlation based techniques on the EEG data to aid the selection of relevant stimulus related components has to be evaluated.

Furthermore applying the proposed method to a simpler event-related paradigm might aid to evaluate the whole pre-processing strategy used on the EEG data and might also help in evaluating the procedure used to select relevant components.

Apart from the study of the coupling between EEG power modulations and fMRI BOLD signal, the proposed approach could be used to create an experiment related lead-field matrix to be used to solve the inverse problem in non simultaneous data of the same experiment. In particular, the matrix that defines the contribution of each voxel in the fMRI to all channels of the EEG (lead-field) could be estimated training multivariate models to predict the single channels EEG modulations and using the resulting fMRI maps.

In summary, multivariate regression based on predictive models appears to be a valuable tool to analyse data obtained from different modalities and to study the coupling between the EEG power modulations and the fMRI BOLD signals.

References

Allen, P.J., Josephs, O., Turner, R. (2000). A method for removing imaging artifact from continuous EEG recorded during functional MRI. Neuroimage; 12(2):230-9.

Bishop CM. Pattern Recognition and Machine Learning. Springer 2006.

Caetano, G., Jousmaki, V., Hari, R. (2007). Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions. Proc. Nat. Acad. Sci.; 104(21): 9058-9062.

Dale, A.M., Halgren, E. (2001). Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. Current opinion in neurobiology; 11:202-208.

Debener, S., Ullsperger, M., Siegel, M., Engel, A.K. (2006). Single-Trial EEG-fMRI reveals the dynamics of cognitive function. Trends in Cognitive Sciences; 10(12): 558-563.

Debener, S., Strobel, A., Sorger, B., Peters, J., Kranczioch, C., Engel, A.K., Goebel, R. (2007). Improved quality of auditory event-related potentials recorded simultaneously with 3-T fMRI: removal of the ballistocardiogram artefact. Neuroimage; 34(2):587-97.

Delorme, A., Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci Methods;134(1):9-21.

de Munck, J.C., Gonçalves, S.I., Huijboom, L., Kuijer, J.P., Pouwels, P.J., Heethaar, R.M., Lopes da Silva, F.H. (2007). The hemodynamic response of the alpha rhythm: an EEG/fMRI study. Neuroimage; 35(3):1142-51.

Duda RO, Hart PE, Stork DG. Pattern Classification. John Wiley & Sons, 2nd ed. 2001.

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., and Di Salle, F., 2005. Independent component analysis of fMRI group studies by self-organizing clustering. NeuroImage 25(1), 193-205.

Feige, B., Scheffler, K., Esposito, F., Di Salle, F., Hennig, J., Seifritz, E. (2005). Cortical and subcortical correlates of electroencephalographic alpha rhythm modulation. J Neurophysiol; 93(5):2864-72.

Formisano, E., De Martino, F., Valente G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using Machine Learning. Magnetic Resonance Imaging; In Press.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., Turner, R. (1998). Event-Related fMRI: Characterizing Differential Responses. Neuroimage; 7(1): 30-40.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J. (2008). Bayesian decoding of brain images. Neuroimage; 39(1):181-205.

Giraud, A.L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S., Laufs, H. (2007). Endogenous Cortical Rhythms Determine Cerebral Specialization for Speech Perception and Production. Neuron; 56(6):1127-1134.

Goldman, R.I., Stern, J.M., Engel, J. Jr, Cohen, M.S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. Neuroreport; 13(18):2487-92.

Gonçalves, S.I., de Munck, J.C., Pouwels, P.J., Schoonhoven, R., Kuijer, J.P., Maurits, N.M., Hoogduin, J.M., Van Someren, E.J., Heethaar, R.M., Lopes da Silva F.H. (2006). Correlating the alpha rhythm to BOLD using simultaneous EEG/fMRI: inter-subject variability. Neuroimage; 30(1):203-13.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V. (2003). Functional connectivity of the resting brain: a network analysis of the default mode hypothesis. Proc. Nat. Acad. Sci.; 100(1):253-8.

Guyon I, Elisseeff A. An introduction to variable and feature selection (2003). Journal of Machine Learning Research; 3:1157–1182.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. Science; 303(5664):1634-40.

Herrmann, C.S., Debener, S. (2007). Simultaneous recording of EEG and BOLD responses: A historical perspective. International Journal of Psychophysiology; doi:10.1016/j.ijpsycho.2007.06.006.

Hyvärinen A, Karhunen J, Oja E. Independent Component Analysis. John Wiley & Sons 2001.

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., Krakow, K. (2003a). EEG-correlated fMRI of human alpha activity. Neuroimage; 19(4):1463-76.

Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., Kleinschmidt, A. (2003b). Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. Proc. Natl. Acad. Sci.; 100(19):11053-8.

Laufs, H., Holt, J.L., Elfont, R., Krams, M., Paul, J.S., Krakow, K., Kleinschmidt, A.

(2006). Where the BOLD signal goes when alpha EEG leaves. Neuroimage; 31(4):1408-18.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. and Oeltermann, A.(2001). Neurophysiological Investigation of the Basis of the fMRI signal. Nature; 412:150-157.

MacKay DJC. Bayesian methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, Models of Neural Networks III, 1994, ch. 6, pages 211–254. Springer-Verlag, New York.

Makeig, S., Jung, T.P., Bell, A.J., Ghahremani, D., Sejnowski, T.J. (1997). Blind separation of auditory event-related brain responses into independent components.

Proc. Natl. Acad. Sci.; 94(20):10979-84.

Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L., Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. Proc. Natl. Acad. Sci.; 104(32):13170-5.

Martínez-Montes, E., Valdés-Sosa, P.A., Miwakeichi, F., Goldman, R.I., Cohen, M.S. (2004). Concurrent EEG/fMRI analysis by multiway Partial Least Squares. Neuroimage; 22(3):1023-34.

Moosmann, M., Ritter, P., Krastel, I., Brink, A., Thees, S., Blankenburg, F., Taskin, B., Obrig, H., Villringer, A. (2003). Correlates of alpha rhythm in functional magnetic resonance imaging and near infrared spectroscopy. Neuroimage; 20(1):145-58.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., Malach1, R. (2005). Coupling between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. Science; 309:951.

Neal, R.M. Bayesian Learning for Neural Networks. Springer 1996.

Niazy, R.K., Beckmann, C.F., Iannetti, G.D., Brady, J.M., Smith, S.M. (2005). Removal of FMRI environment artifacts from EEG data using optimal basis sets. Neuroimage; 28(3):720-37.

Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press 2006.

Ritter, P., Villringer, A. (2006). Simultaneous EEG-fMRI. Neuroscience and behavioral reviews; 30:823-838.

Salmelin, R., Hari, R. (1994). Characterization of spontaneous MEG rhythms in healthy adults. Electroencephalography and Clinical Neurophysiology; 91:237-248.

Scheeringa, R., Bastiaansen, M.C., Petersson, K.M., Oostenveld, R., Norris, D.G., Hagoort, P. (2007). Frontal theta EEG activity correlates negatively with the default mode network in resting state. International Journal of Psychophysiology; doi: 10.1016/j.ijpsycho.2007.05.017.

Tipping ME Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 2001; 1:211–244.

Vapnik, V.N. The nature of statistical learning theory. Springer-Verlag New York 1995.