Combined Permutation Test and Mixed-Effect Model for Group Average Analysis in fMRI

Sébastien Mériaux,[1,2]* Alexis Roche,[1,2] Ghislaine Dehaene-Lambertz,[2,3] Bertrand Thirion,[4] and Jean-Baptiste Poline[1,2]

[1]*CEA, Service Hospitalier Frédéric Joliot, Orsay, France*
[2]*IFR 49, Institut d'Imagerie Neurofonctionnelle, Paris, France*
[3]*INSERM U 562, Service Hospitalier Frédéric Joliot, Orsay, France*
[4]*INRIA Futurs, Orsay, France*

Abstract: In group average analyses, we generalize the classical one-sample t test to account for heterogeneous within-subject uncertainties associated with the estimated effects. Our test statistic is defined as the maximum likelihood ratio corresponding to a Gaussian mixed-effect model. The test's significance level is calibrated using the same sign permutation framework as in Holmes et al., allowing for exact specificity control under a mild symmetry assumption about the subjects' distribution. Because our likelihood ratio test does not rely on homoscedasticity, it is potentially more sensitive than both the standard t test and its permutation-based version. We present results from the Functional Imaging Analysis Contest 2005 dataset to support this claim. *Hum Brain Mapp 27:402–410, 2006.*

Key words: group analysis; mixed effects; permutation test; fMRI

INTRODUCTION

To date, the vast majority of inference procedures available in neuroimaging analysis packages rely on standard t or F decision statistics. Underlying those statistics is an implicit homogeneity assumption about the input data; more precisely, the measurements are assumed identically and normally distributed (conditionally on the model parameters). In single-subject analyses, this is reflected by the common usage of stationary Gaussian noise models for functional MRI (fMRI) time series. Although idealistic, such models may produce reliable inferences provided the degrees of freedom (DF) are large enough, i.e., the number of unknown parameters involved is small compared to the number of scans acquired for one subject.

However, in group random-effect analyses, where the goal is to generalize single-subject findings to a population, DF are limited by the usually small number of subjects undergoing the same fMRI experiment. In this case, even subtle deviations from homoscedasticity or normality may result in significantly biased inferences, in terms of both specificity (control of false-positives) and sensitivity (control of false-negatives). Previous studies [Kherif et al., 2004; Mériaux et al., 2004] have revealed that the homogeneity assumption is often violated in fMRI group datasets.

Permutation testing methods [Brammer et al., 1997; Bullmore et al., 1999; Hayasaka and Nichols, 2003; Holmes et al., 1996; Nichols and Holmes, 2002] have the potential to correct for specificity bias, as they can calibrate any test based on mild, nonparametric distributional assumptions. In addition, because they circumvent the classical approximations

Contract grant sponsor: French Ministry of Research (ACI Neurosciences Intégrées et Computationnelles and ACI Masses de Données).

*Correspondence to: Sébastien Mériaux, CEA, Service Hospitalier Frédéric Joliot, Orsay, France. E-mail: meriaux@shfj.cea.fr

Received for publication 20 October 2005; Accepted 27 December 2005

DOI: 10.1002/hbm.20251

Published online 4 April 2006 in Wiley InterScience (www.interscience.wiley.com).

underlying random field theory (RFT), they offer a robust solution to the multiple comparison problem. However, currently available permutation tests, as implemented for instance in the Statistical nonParametric Mapping (SnPM) toolbox, use standard *t* or *F* statistics, and may henceforth be suboptimally sensitive.

Over the past few years, "mixed-effect" models have been proposed [Beckmann et al., 2003; Friston et al., 2002; Neumann and Lohmann, 2003; Woolrich et al., 2004; Worsley et al., 2002] in order to relax the usual assumption that BOLD effects estimated from the within-subject analysis level are identically distributed across subjects. Their ground motivation is that "fixed-effect" variances (the within-subject variances estimated in the first-level analyses) are bound to be subject-dependent, reflecting the simple fact that some subjects may yield more accurate response estimates than others.

While mixed-effect models account for possibly heterogeneous first-level variances, their statistical calibration is usually performed based on the assumption that the random effects (only estimates of which are available) are normally distributed across subjects. We propose here to fill the gap between mixed-effect models and permutation testing approaches, thus designing a permutation test that employs a "mixed-effect" decision statistic. More precisely, our decision statistic is defined as the maximum likelihood ratio corresponding to a Gaussian mixed-effect model. It amounts to a nonstandard *t* statistic, which essentially reweights the subjects in a nonuniform fashion according to the reliability of their respective estimated effects.

MATERIALS AND METHODS

Assume *n* subjects were scanned during a cognitive experiment, and their respective fMRI data were processed individually so that, for each subject *i*, a pair of summary statistics $(\hat{\beta}_i, \hat{\sigma}_i)$is available: $\hat{\beta}_i$ is an image of estimated BOLD effects relative to a given contrast of experimental conditions and $\hat{\sigma}_i$ is an image of voxelwise standard error estimates of $\hat{\beta}_i$. For the sake of clarity, we will restrict ourselves to scalar (one-dimensional) effects in this article. Our aim is to perform a random-effect analysis on the mean population effect using the image pairs $(\hat{\beta}_i, \hat{\sigma}_i)$ as input data. Although generally not exhaustive for the unknown *true* subject's effect β_i(in particular, we do not retain DF and spatial covariance information from the first-level analysis), those summary statistics exploit more information from the fMRI data than the effect images alone, with the potential to produce more sensitive inferences.

Nonparametric Model

In order to relate the individual summary statistics to the unknown population mean effect β_G, we adopt a nonparametric two-level model:

$$\begin{cases} (\hat{\beta}_i, \hat{\sigma}_i)|\beta_i \sim f_i(\hat{\beta}_i, \hat{\sigma}_i|\beta_i) \\ \beta_i|\beta_G \sim g(\beta_i|\beta_G), \end{cases} \tag{1}$$

where the conditional density f_i models the within-subject variability inherent to any statistical approach to fMRI signal modeling. We allow f_i to be subject-dependent in order to account for heteroscedasticity, i.e., possibly different amounts of noise across datasets. At the second level, the density *g* models the intrinsic between-subject variability of the BOLD response. Marginalizing out the true effect β_i, the two-level model may be compacted into:

$$p_i(\hat{\beta}_i, \hat{\sigma}_i|\beta_G) = \int f_i(\hat{\beta}_i, \hat{\sigma}_i|\beta_i) g(\beta_i|\beta_G) d\beta_i, \tag{2}$$

which accounts for the composite variability resulting from both within-subject and between-subject sources of randomness.

In the remainder of this article we will work under the following statistical assumptions:

(A_1) The input statistic images are independently, although nonidentically, distributed so that their joint distribution has a factored form:

$$p(\hat{\beta}_1, \hat{\sigma}_1, \hat{\beta}_2, \hat{\sigma}_2, \ldots, \hat{\beta}_n, \hat{\sigma}_n|\beta_G) = \prod_{i=1}^{n} p_i(\hat{\beta}_i, \hat{\sigma}_i|\beta_G) \tag{3}$$

(A_2) The true effect is symmetrically distributed in the population of interest, i.e., the density $g(\beta_i|\beta_G)$ is symmetric with respect to β_G.

(A_3) First-level estimators are location equivariant and scale invariant:

$$\forall\, i,\ \forall (a, b) \in \mathbb{R}^2, \quad f_i(\hat{\beta}_i, \hat{\sigma}_i|\beta_i) = |a| f_i(a\hat{\beta}_i + b, |a|\hat{\sigma}_i|a\beta_i + b) \tag{4}$$

Assumption (A_1) is justified as long as subjects are drawn independently from the population to which findings are to be generalized, and BOLD signal measurement errors induced by the scanner are not reproducible across sessions. Population symmetry (A_2) is the usual assumption underlying permutation tests in group average analyses [Holmes et al., 1996; Nichols and Holmes, 2002], and is milder than normality. Finally, (A_3) is a natural consistency requirement that is met by standard within-subject estimation techniques based on the general linear model. Together (A_2) and (A_3) imply that each subject's summary statistics are symmetrically distributed about the mean population effect, that is: $p_i(2\beta_G - \hat{\beta}_i, \hat{\sigma}_i|\beta_G) = p_i(\hat{\beta}_i, \hat{\sigma}_i|\beta_G)$.

Permutation Test Framework

We wish to test the global null hypothesis H_0: $\beta_G = 0$ that all voxels have a zero mean population effect. For each voxel *k*, consider a decision statistic:

$$D_k \equiv d\Big((\hat{\beta}_{1k}, \hat{\sigma}_{1k}), (\hat{\beta}_{2k}, \hat{\sigma}_{2k}), \ldots, (\hat{\beta}_{nk}, \hat{\sigma}_{nk})\Big), \tag{5}$$

that depends on the input data at voxel k regardless of other voxels. This property guarantees the subset pivotality condition required for family-wise error strong control [Nichols and Hayasaka, 2003]. Notice that statistics using locally pooled variance estimators [Holmes et al., 1996; Worsley et al., 2002] require additional assumptions to fulfill subset pivotality.

Assumptions (A_1), (A_2), and (A_3) stated above imply that estimated effects' signs are exchangeable under H_0, that is: each subject's data can be arbitrarily shuffled according to $(\hat{\beta}_i, \hat{\sigma}_i) \rightarrow (-\hat{\beta}_i, \hat{\sigma}_i)$ without modifying the joint distribution of all subjects' data. This exchangeability property yields a straightforward generalization of the permutation framework developed in Holmes et al. [1996] and Nichols and Holmes [2002] where first-level variances can now be taken into account.

The permutation test consists of tabulating the multivariate null distribution of the decision statistics $(D_1,\ldots,D_K)$ by permuting the first-level statistic images across all possible flips of effect images' signs, the number of which is 2^n. The resulting distribution is, in fact, conditional on the nonexchangeable part of the observations, namely, the absolute effect and standard error images $(|\hat{\beta}_i|, \hat{\sigma}_i)$. In this conditional sense, the permutation test is exact, although its sensitivity is intrinsically limited by the finite number of possible permutations.

Single threshold test

We use critical regions $D_k \geq \delta$ to threshold the decision statistical map, where δ is constant across the search volume and is chosen so as to control the false-positive rate (FPR) at a specified level α. Hence, we have to solve:

$$\alpha = \mathbb{E}(FP|H_0) = \frac{1}{K}\sum_k P(D_k \geq \delta|H_0), \tag{6}$$

where K is the number of voxels in the search volume and $P(D_k \geq \delta|\mathrm{H}_0)$ is the marginal reference probability that D_k exceeds the threshold, which is a priori voxel-dependent unless the decision map is stationary. Exploiting subset pivotality, each distribution $p_k(D|H_0)$ may be tabulated by permuting voxelwise data only. From Eq. (6), δ is then found to be the $100(1 - \alpha)$-th percentile of the across-voxel average distribution, $\bar{p}(D) \equiv (\frac{1}{K}) \sum_k p_k(D|H_0)$. In practice, we approximate $\bar{p}$ over a sufficient number of randomly selected voxels in order to save both computation time and memory load.

Once the FPR-controlling threshold is tuned, P values corrected for multiple comparisons are computed using an imagewise permutation test on the $D_{\max}$ statistic, similar to Holmes et al. [1996] and Nichols and Holmes [2002]. The test also tabulates the distribution of the maximum suprathreshold cluster size [Bullmore et al., 1999; Hayasaka and Nichols, 2003], hence producing cluster-level P values for each FPR-surviving cluster. This second stage is typically more time-consuming.

Decision Statistic

An essential component of the test procedure is the decision statistic D_k, which, for optimal sensitivity, should be chosen according to the Neyman-Pearson theorem, provided that the statistical model is fully specified. However, being nonparametric (see Nonparametric Model), our model is obviously misspecified. A natural workaround is then to consider a parametric restriction of the model and derive the decision statistic accordingly. We shall stress that restrictive assumptions do not need to hold for the test's specificity to be correctly assessed; however, a potential lack of sensitivity is to be expected if the true distribution of the data deviates substantially from the restricted model.

Following previous work on hierarchical linear models [Beckmann et al., 2003; Friston et al., 2002; Neumann and Lohmann, 2003; Woolrich et al., 2004; Worsley et al., 2002], our restricted model assumes locally normal distributions both at the within-subject and between-subject levels:

$$\begin{cases} f_i(\hat{\beta}_{ik}, \hat{\sigma}_{ik}|\beta_{ik}) = \psi\left(\dfrac{\hat{\beta}_{ik} - \beta_{ik}}{\hat{\sigma}_{ik}}\right) f_i(\hat{\sigma}_{ik}) \\ g(\beta_{ik}|\beta_{Gk}) = \psi\left(\dfrac{\beta_{ik} - \beta_{Gk}}{\sigma_{Gk}}\right), \end{cases} \tag{7}$$

where ψ denotes the normalized Gaussian, and $f_i(\hat{\sigma}_{ik})$ is an arbitrary marginal distribution assumed independent from β_{ik}, which will hence play no role in the decision statistic. This restricted model is specified up to only one hyperparameter, namely, the group standard deviation σ_{Gk}. Notice here within-subject variance estimates are implicitly assimilated with their true values. Sensitivity could be further improved by considering the effective DF, hence using Student distributions instead of Gaussians at the first level. This approach, however, leads to significantly increased computation time.

As is a customary frequentist approach, we define the decision statistic as the (log) maximum likelihood ratio (MLR):

$$D_k \equiv -2\log\left[\frac{\sup\limits_{\sigma_{Gk}\in\mathbb{R}_+} L(0, \sigma_{Gk})}{\sup\limits_{(\beta_{Gk}, \sigma_{Gk})\in\mathbb{R}\times\mathbb{R}_+} L(\beta_{Gk}, \sigma_{Gk})}\right], \tag{8}$$

where $L(\beta_{Gk}, \sigma_{Gk})$ is the likelihood function associated with the restricted model, which reads:

$$L(\beta_{Gk}, \sigma_{Gk}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma_{Gk}^2 + \hat{\sigma}_{ik}^2}} \exp\left[-\frac{(\hat{\beta}_{ik} - \beta_{Gk})^2}{2(\sigma_{Gk}^2 + \hat{\sigma}_{ik}^2)}\right] \tag{9}$$

Essentially, the MLR compares the profile likelihood of H_0 with the maximum profile likelihood over all alternative hypotheses. It is easily seen that the ratio in Eq. (8) ranges from 0 to 1, hence D_k is always nonnegative. Since a high value of D_k indicates that H_0 is unlikely, the critical region for the likelihood ratio test is of the form $D_k \geq \delta$, where δ is some threshold. Owing to Wilks' theorem, D_k is actually an approximate χ_1^2-score. This result provides a rough significance assessment, but is of little use to accurately control specificity because the χ^2 approximation is only valid asymptotically (for large samples) and under the restricted model.

Computation

Except in special cases (see Consistency with Student's statistic), none of the two maximum likelihood problems involved in Eq. (8) can be solved explicitly. In practice, numerical solutions are found using an expectation-maximization (EM) algorithm [Dempster et al., 1977] detailed in the Appendix. Convergence towards a unique solution is always guaranteed because the likelihood function (9) can be shown to have a single global maximum, both in the constrained (fixed $\beta_G = 0$) and unconstrained settings.

One-sided test statistic

So far we have considered the point-null hypothesis, H_0: $\beta_G = 0$, upon rejection of which it is impossible to conclude about the mean effect's sign. This is the typical drawback of two-sided tests such as the F test as compared to one-sided tests such as the t test. To derive a decision statistic suitable for one-sided testing, we need to recast H_0 as the composite hypothesis H_0: $\beta_G \leq 0$, and redefine the MLR statistic accordingly. The resulting MLR boils down to a simple sign modulation of the previously defined MLR (8):

$$\tilde{D}_k = \mathrm{sign}(\hat{\beta}_{Gk})\,\sqrt{D_k}, \qquad (10)$$

where $\hat{\beta}_{Gk}$ is the maximum likelihood estimate of β_{Gk}, as provided by the EM algorithm. Square rooting is used to render $\tilde{D}_k$ roughly comparable with a z-score (see Decision Statistic). The one-sided test can be calibrated using the sign permutation framework described in Permutation Test Framework (above), although a monotonicity property is required. For any subset of voxels ν, any threshold δ and any uniformly nonpositive map β_G ($\forall k \in \mathcal{V}$, $\beta_{Gk} \leq 0$), the decision statistic needs to verify:

$$P(\cup_{k\in\mathcal{V}}\{\tilde{D}_k \geq \delta\}|\beta_G) \leq P(\cup_{k\in\mathcal{V}}\{\tilde{D}_k \geq \delta\}|H_0) \qquad (11)$$

It turns out that the property holds for $\tilde{D}_k$ (the proof mainly relies on the fact that $\tilde{D}_k$ is decreasing under negative shifts in the samples).

Consistency with Student's statistic

In the case of homoscedasticity, that is, when all first-level variances are identical ($\hat{\sigma}_{1k} = \hat{\sigma}_{2k} = \ldots = \hat{\sigma}_{nk}$), the one-sided MLR enjoys a closed-form expression which is a strictly increasing function of the standard t statistic:

$$\tilde{D}_k = \mathrm{sign}(T_k)\sqrt{n\log\left(1 + \frac{T_k^2}{n-1}\right)},$$

$$\text{with } T_k \equiv \frac{\sqrt{n(n-1)}\,\hat{\beta}_{Gk}}{\sqrt{\Sigma_i(\hat{\beta}_{ik} - \hat{\beta}_{Gk})^2}}, \qquad (12)$$

given that, in this special instance, $\hat{\beta}_{Gk}$ identifies with the classical sample mean. This implies that, under homoscedastic measurements, the one-sided MLR is fully equivalent to the t statistic from a decision theoretical standpoint. This justifies referring to $\tilde{D}_k$ as a mixed-effect t statistic.

We note that Worsley et al. [2002] proposed a similar extension of the t statistic, which turns out to be an MLR variant in which the nuisance parameter δ_{Gk} is preestimated by restricted maximum likelihood, then held fixed in Eq. (8) instead of being optimized over two different spaces. We have not carefully studied the practical difference between these approaches.

EXPERIMENTS

We now present results of the method on the Functional Imaging Analysis Contest (FIAC) 2005 dataset described in Dehaene-Lambertz et al. [2006]. While the experimental procedure comprised four sessions, two using a block design and two using an event-related design, only the results of the block experiment are reported here.

Data Processing

First-level analyses were conducted using SPM2 (Statistical Parametric Mapping software). Data were submitted successively to motion correction, slice timing, normalization to the MNI template, and spatial smoothing using a 5 × 5 × 5 mm^3 full-width at half-maximum (FWHM) Gaussian filter. For each of the 15 available subjects, summary statistics were obtained from a fixed-effect analysis on both sessions, except for subject FIAC10, who was known to be sleeping during his second session.

A model consisting of five conditions was set up: Same Sentences – Same Speakers (SSt-SSp), Same Sentences – Different Speakers (SSt-DSp), Different Sentences – Same Speakers (DSt-SSp), and Different Sentences – Different Speakers (DSt-DSp). The first sentence of each block was excluded from all four above conditions and modeled as a fifth condition. Low-frequency drifts were compensated using a temporal highpass filter with a 1/128-Hz cutoff frequency, and noise was modeled as an AR(1) process. Each condition was modeled as a single event; however, further experiments (not reported here) revealed that group analyses were almost insensitive to block duration modeling.

The permutation testing framework described in Materials and Methods was implemented in C and binded with Matlab to be part of the Distance toolbox for SPM (v. beta 2.0, freely available at http://www.madic.org) designed to provide advanced diagnosis and inference tools for group analysis in fMRI. Two particular permutation tests were considered in this study: the one based on the mixed-effect

TABLE I. Results of the contrasts of interest relating to the sentence factor

Cluster anatomical location and statistical test procedure	Cluster-level, P_{corr}	Cluster extent (voxels)	Voxel-level peak, P_{corr}	Peak position (mm): x, y, z
Sentence effect				
Left superior temporal sulcus (middle STS)				
Parametric *t* test (SPM)	**0.00** (**0.00**)	255 (275)	0.00 (0.01)	−63, −15, 0
Permutation *t* test	**0.00** (**0.03**)	264 (285)	0.00 (0.01)	−63, −15, 0
Permutation MFX *t* test	**0.00** (**0.03**)	322 (378)	0.00 (0.00)	−63, −15, 0
Right superior temporal sulcus (anterior STS)				
Parametric *t* test (SPM)	**0.03** (0.16)	49 (49)	0.12 (0.99)	60, 0, −3
Permutation *t* test	**0.05** (0.35)	51 (51)	0.03 (0.45)	60, 0, −3
Permutation MFX *t* test	**0.05** (0.33)	60 (72)	0.00 (0.00)	63, −3, 0
DStSSp > SStSSp				
Left superior temporal sulcus (middle STS)				
Parametric *t* test (SPM)	**0.00** (**0.00**)	247 (261)	0.02 (0.24)	−63, −12, 3
Permutation *t* test	**0.00** (**0.04**)	259 (277)	0.01 (0.09)	−63, −12, 3
Permutation MFX *t* test	**0.00** (**0.04**)	333 (377)	0.00 (0.00)	−63, −15, 0
Right superior temporal sulcus (anterior STS)				
Parametric *t* test (SPM)	**0.00** (**0.02**)	74 (75)	0.12 (0.99)	60, −6, −3
Permutation *t* test	**0.01** (0.16)	84 (85)	0.04 (0.41)	60, −6, −3
Permutation MFX *t* test	**0.01** (0.12)	147 (169)	0.01 (0.02)	60, −6, −3
DStDSp > StDSp				
Left superior temporal sulcus (middle STS)				
Parametric *t* test (SPM)	**0.00** (**0.00**)	112 (113)	0.00 (0.03)	−63, −15, 0
Permutation *t* test	**0.01** (0.10)	115 (116)	0.00 (0.02)	−63, −15, 0
Permutation MFX *t* test	**0.01** (0.10)	146 (164)	0.00 (0.00)	−63, −15, 0
Left superior temporal sulcus (posterior STS)				
Parametric *t* test (SPM)	**0.02** (**0.02**)	54 (66)	0.74 (1.00)	−51, −39, 3
Permutation *t* test	**0.04** (0.23)	58 (70)	0.32 (0.95)	−51, −39, 3
Permutation MFX *t* test	**0.04** (0.25)	72 (92)	0.04 (0.17)	−51, −39, 3

Whole-brain analysis results in parentheses.

(MFX) *t* statistic (see Decision Statistic) and the one based on the standard *t* statistic. Notice that the latter is equivalent to the one-sample test implemented in SnPM when no variance smoothing is selected.

For all contrasts presented here, the statistical maps were thresholded at $P \leq 0.01$ uncorrected, then corrected for cluster extent at 5%. Practically, the height threshold was computed using approximate voxelwise permutation tests involving 2,000 random permutations instead of $2^{15} = 32{,}768$ exhaustive permutations, and then averaging reference distributions over 1,000 randomly selected voxels (see Computation). Cluster-size inference was carried out using an approximate permutation test involving 10,000 random permutations. This implies that cluster-level *P* values are approximated with a standard error of $\surd(P - P^2/10{,}000)$, and are thus accurate to the second decimal point. Clusters were defined in the sense of the 18-connectivity.

We performed group analyses on a manually segmented symmetrical mask of 2,920 voxels surrounding the perisylvian areas, which are known to be linguistically and acoustically sensitive and are the ones that the experimental protocol was intended to study. However, for comparison with other results published in this special issue, we also report whole-brain analyses (45,484 voxels) that are inevitably more conservative in terms of corrected inferences (results are given in parentheses in the tables below). For each mask-restricted analysis the total computation time was on the order of 20 seconds for the permutation *t* test and 15 minutes for the permutation MFX *t* test on a standard PC (2.80-GHz single processor) running Linux. These times were significantly higher, respectively, about 5 min and 3 h 45 min, for the whole-brain analyses.

We also report results of the parametric *t* test as implemented in SPM, when thresholded at the same level ($P \leq 0.01$ uncorrected) using the Student distribution with 14 DF. Cluster-size inference in SPM uses an approximation formula based on random field theory [Friston et al., 1994]. In the following, clusters are reported if their corrected *P* value is less than 5% in at least one of the three statistical procedures. We do not account for the multiple comparison problem associated with performing several tests on several contrasts, as the scope of this article is to compare analysis methods under various conditions.

RESULTS

We first report the group analysis results obtained considering each factor separately while holding the other factor fixed or not (sentence in the following section, then speaker in the Speaker Factor section). The results of the interaction between the two factors are presented in the section Interaction.

Sentence Factor

Table I summarizes the results of group average activation obtained when focusing on the sentence factor. It shows

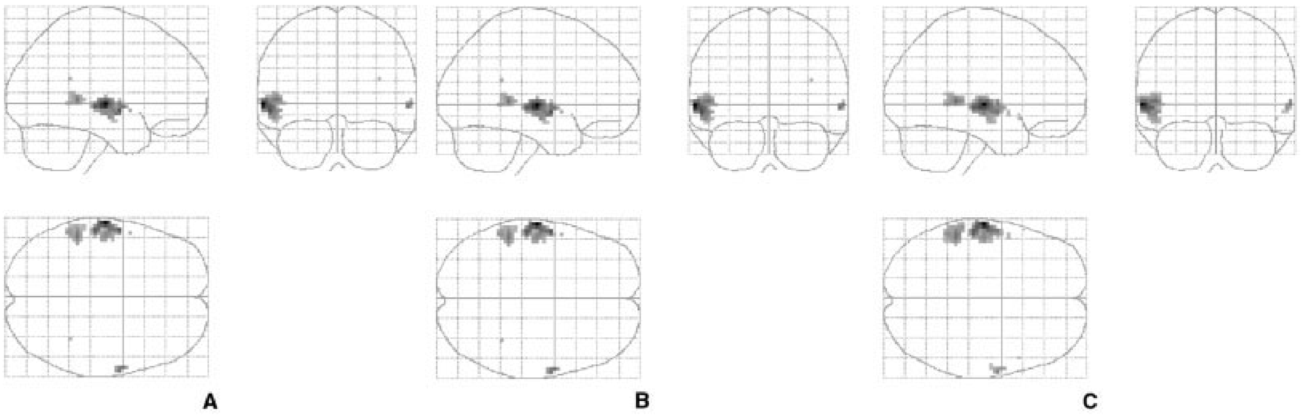

Figure 1.
Maximum intensity projection (MIP) of group average activation maps obtained for the DStDSp > SStDSp contrast (no cluster-level correction). See Table I. **A:** Parametric *t* test (SPM). **B:** Permutation *t* test. **C:** Permutation MFX *t* test.

very similar suprathreshold clusters for the three compared statistical procedures (Fig. 1). We notice that the SPM approach is the least conservative in terms of cluster-level *P* values, although the permutation MFX *t* test approach detects the most extended cluster. This issue will be addressed in the Discussion.

The sentence main effect (DStSSp + DStDSp > SStSSp + SStDSp), reveals two bilateral suprathreshold clusters (first two clusters in Table I): one widespread cluster located in the left STS middle and one less extended in the right STS anterior. As surveyed in Belin et al. [2004], several studies have outlined the same specific involvement in speech perception of the middle and anterior parts of the bilateral STS.

When restricting the sentence main effect to the same speaker (i.e., DStSSp > SStSSp), the same clusters (third and fourth clusters in Table I) are found in the left and right STS. Moreover, when restricting the sentence main effect to different speakers (DStDSp > SStDSp), the right STS cluster does not remain significant at the considered thresholds, while the left STS cluster seems to subdivide in one middle STS cluster (fifth cluster in Table I) and another smaller posterior STS cluster (sixth cluster in Table I).

These results indicate that the left STS (Fig. 1C) may be seen as a normalization region for the speaker factor, as it responds identically to sentence variations whether the speaker changes or remains the same. This region seems to respond significantly to linguistic information, which is in accordance with Pallier et al. [2003] and Dehaene et al. [1997].

Speaker Factor

Table II summarizes the results obtained when focusing on the speaker factor. The only clusters surviving 5% cluster extent correction are detected using the permutation MFX *t* test approach, confirming the benefit of taking into account the first-level variability. It is also interesting to notice the disagreement between the corrected cluster-level *P* values as

TABLE II. Results of the contrasts of interest relating to the speaker factor

Cluster anatomical location and statistical test procedure	Cluster-level, P_{corr}	Cluster extent (voxels)	Voxel-level peak, P_{corr}	Peak position (mm): *x*, *y*, *z*
Speaker effect				
Left superior temporal sulcus (posterior STS)				
Parametric *t* test (SPM)	0.53 (1.00)	14 (14)	0.62 (1.00)	−63, −42, 9
Permutation *t* test	0.18 (0.89)	14 (14)	0.19 (0.94)	−63, −42, 9
Permutation MFX *t* test	**0.05** (0.35)	51 (59)	0.07 (0.36)	−63, −42, 9
DSpSSt > SSpSSt				
Left superior temporal sulcus (posterior STS)				
Parametric *t* test (SPM)	0.13 (0.64)	30 (30)	0.82 (1.00)	−66, −45, 9
Permutation *t* test	0.07 (0.45)	33 (33)	0.35 (0.99)	−66, −45, 9
Permutation MFX *t* test	**0.03** (0.21)	83 (103)	0.07 (0.32)	−60, −30, 6
Right superior temporal sulcus (middle STS)				
Parametric *t* test (SPM)	0.41 (0.99)	17 (17)	0.73 (1.00)	57, −12, −3
Permutation *t* test	0.15 (0.75)	18 (18)	0.28 (0.96)	57, −12, −3
Permutation MFX *t* test	**0.04** (0.25)	68 (84)	0.12 (0.49)	57, −12, −3

Whole-brain analysis results in parentheses.

TABLE III. Results of the *Sentence* × *Speaker* interaction

Cluster anatomical location	Statistical test procedure	Cluster-level P_{corr}	Cluster extent (voxels)	Voxel-level peak P_{corr}	Peak position (mm): x, y, z
Right superior	Parametric *t*-test (SPM)	0.13 (0.82)	24 (24)	0.70 (1.00)	60, −12, −3
Temporal sulcus	Permutation *t*-test	0.09 (0.64)	25 (25)	0.28 (0.94)	60, −12, −3
(Middle STS)	Permutation MFX *t*-test	**0.02** (0.20)	97 (97)	0.04 (0.28)	60, −12, −3

Whole-brain analysis results in parentheses.

estimated by SPM and by the permutation *t* test, although the cluster extents are very similar. This suggests that RFT approximations are not valid here (see Discussion).

The speaker main effect (SStDSp + DStDSp > SStSSp + DStSSp) yields only one significant cluster located in the left STS posterior (first cluster in Table II). The same cluster pops up when restricting the speaker effect to the same sentence (i.e., DSpSSt > SSpSSt) (second cluster in Table II). For this contrast of interest, the MFX *t* test approach is able to detect another significant cluster located in the right STS middle (third cluster in Table II). Finally, no significant cluster is detected when restricting the speaker effect to different sentences (DSpDSt > SSpDSt).

These results indicate that the left STS posterior region previously identified as responding to sentences (see Sentence Factor) also keeps acoustic information and is not selectively sensitive to the linguistic content of speech.

Interaction

To understand the role of the detected right STS region, we also investigated the interaction between the two factors, defined as DStSSp + SStDSp > SStSSp + DStSSp. Table III summarizes the results of group average activation obtained for this contrast of interest. Again, the permutation MFX *t*-test clearly appears as the most sensitive method and reveals the same right STS middle cluster as in DSpSSt > SSpSSt (Fig. 2).

These results indicate that the right middle STS may not encode for repetition of linguistic aspects of voice when acoustic variation is introduced. As reported in previous studies [Belin et al., 2000; Belin and Zatorre, 2003; von Kriegstein et al., 2003], this region might be involved in paralinguistic aspects of voice processing, such as speaker variation, as evidenced in this particular study.

DISCUSSION

The main purpose of our study was to demonstrate the benefit of combining permutations with a mixed-effect decision statistic. It is interesting to notice that the gain in sensitivity (as compared to the permutation *t* test) is moderate when the group mean effect dominates the first-level standard errors. This situation is well illustrated on the sentence effect (see Sentence Factor), for which effects were all significant at the within-subject level. With no surprise, the three group-level testing procedures produced similar activation maps.

In contrast, when investigating subtle effects that are almost swamped in the first-level variability, a much better sensitivity is achieved using the permutation MFX *t* test. For instance, the latter was the only method to detect a significant cluster in the right STS on the speaker effect (see Speaker Factor). Figure 3 illustrates the Sentence × Speaker interaction in one particular right STS voxel: the heterogeneity observed in the first-level variances, and their high

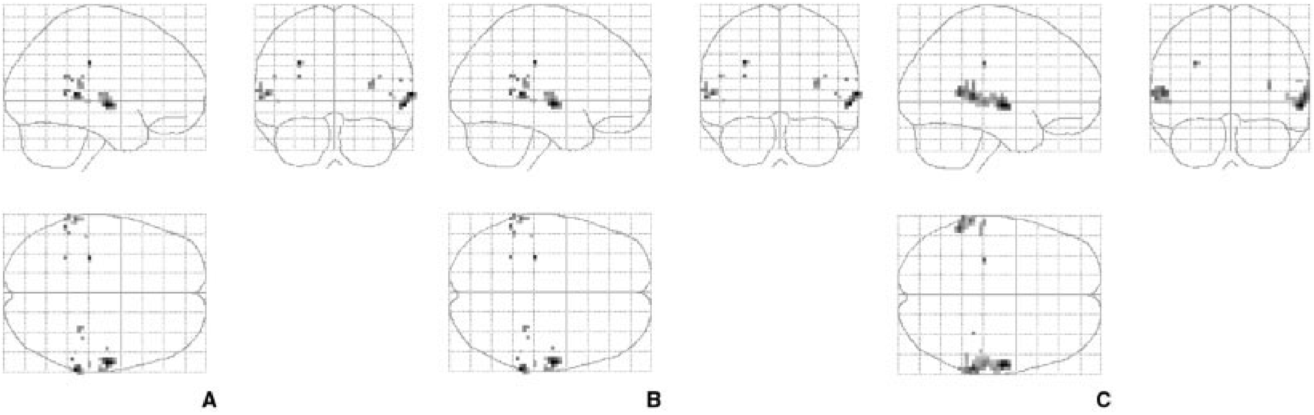

Figure 2.
Maximum intensity projection (MIP) of group average activation maps obtained for the Sentence × Speaker interaction (no cluster-level correction). See Table III. **A:** Parametric *t* test (SPM). **B:** Permutation *t* test. **C:** Permutation MFX *t* test.

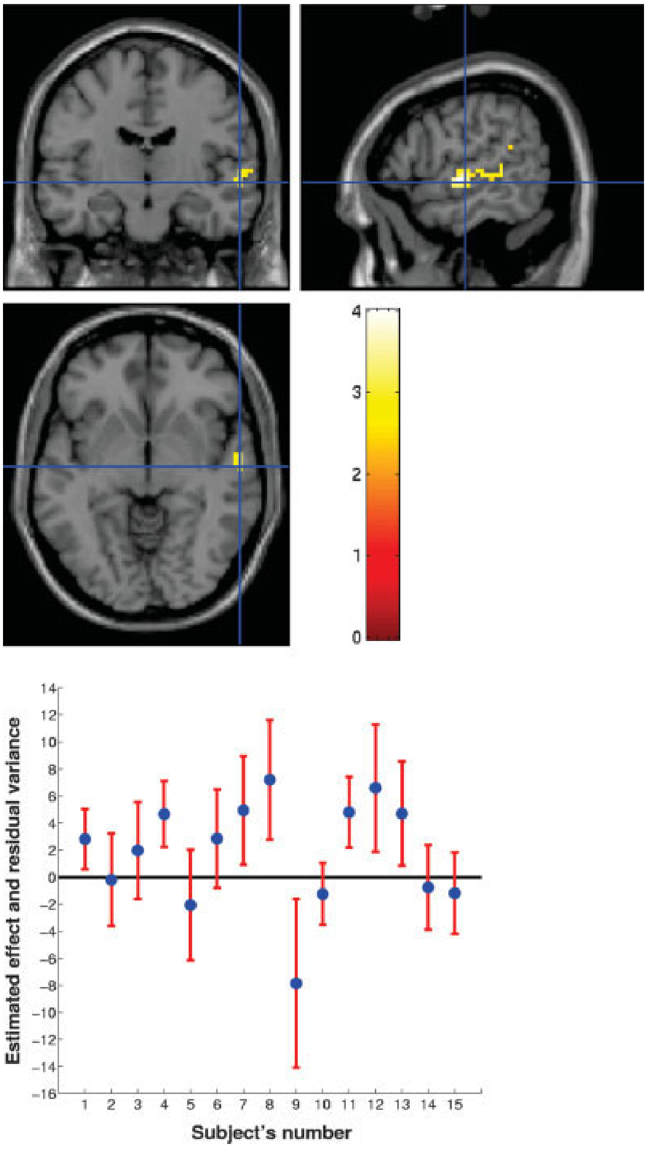

Figure 3.
Sentence × Speaker interaction. (**A**) Group average activation using the permutation MFX *t* test with cross-hair at [60, −15, −6] Talairach coordinates in mm. (**B**) Plot of the estimated effects in the same voxel and associated 70% confidence intervals (within one standard error).

values compared to the estimated effects, explain why a mixed-effect model is necessary to improve detection. The gain in sensitivity is perhaps more spectacular at the voxel level than at the cluster level, presumably due to the small spatial smoothing effect induced by the MFX statistic as it compensates within-subject variabilities.

Also, the experiments enable us to compare two tests that use the same decision statistic: the classical parametric *t* test based on RFT and its permutation version first developed by Nichols and Holmes [2002], where no RFT is invoked for multiple comparison correction. The respective statistical maps have similar height thresholds at the voxel level, indicating that the voxelwise permutation-tabulated cumulative tails are reasonably well approximated by the Student distribution at $P \leq 0.01$.

However, a major disagreement between the corrected P values is observed, both at the voxel level and at the cluster level. In the FIAC study, SPM is generally overconservative. This can be explained by the fact that, on the one hand, the fMRI datasets were smoothed using a moderate 5-mm FWHM and, on the other hand, the group average maps were submitted to a rather low height threshold ($P \leq 0.01$). Those values may well fall outside the validity domain of the RFT approximations underlying SPM [Worsley et al., 1996]. This confirms the wider applicability of permutation approaches to both family-wise error control [Holmes et al., 1996; Nichols and Holmes, 2002] and cluster-size inference [Bullmore et al., 1999; Hayasaka and Nichols, 2003].

From a cognitive point of view, our results confirm previous studies investigating voice-selective areas: the left STS responds to the linguistic content of speech, while the right STS seems to be more sensitive to vocal variations that might subserve speaker identification or speaker emotion.

CONCLUSION

We developed a new one-sample permutation test using a mixed-effect variant of the *t*-statistic. The FIAC dataset reveals significant sensitivity improvement when using the proposed test as compared to both SPM and SnPM one-sample tests. By the time we implemented the method, we observed the same tendency in four datasets out of five. We therefore believe that the method is of practical interest to the neuroimaging community.

The permutation test allows for exact specificity control under a symmetry assumption regarding the distribution of the random effects, in addition to other mild assumptions (see Nonparametric Model). We are not aware of any work confirming or invalidating population symmetry in fMRI group analyses. However, simple examples such as motor cortex activation in both right-handed and left-handed subjects suggest that symmetry does not hold in every circumstance. The sign permutation method would then be inexact. We are currently investigating other resampling schemes, such as the Bootstrap, in order to relax population symmetry.

REFERENCES

Beckmann C, Jenkinson M, Smith S (2003): General multi-level linear modelling for group analysis in fMRI. Neuroimage 20: 1052–1063.

Belin P, Zatorre R (2003): Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14:2105–2109.

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

Belin P, Fecteau S, Bédard C (2004): Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8:129–135.

Brammer M, Bullmore E, Simmons A, Grasby P, Howard R, Woodruff P, Rabe-Hesketh S (1997): Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. Magn Reson Imaging 15:763–770.

Bullmore E, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer M (1999): Global, voxel, and cluster tests, by theory and permutation, for difference between two groups of structural MR images of the brain. IEEE Trans Med Imaging 18:32–42.

Dehaene S, Dupoux E, Mehler J, Cohen L, Paulesu E, Perani D, van de Moortele P, Lehéricy S, LeBihan D (1997): Anatomical variability in the cortical representation of first and second languages. Neuroreport 8:3809–3815.

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

Dempster A, Laird A, Rubin D (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38.

Friston K, Worsley K, Frackowiak R, Maziotta J, Evans A (1994): Assessing the significance of focal activations using their spatial extent. Hum Brain Mapp 1:214–220.

Friston K, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J (2002): Classical and Bayesian inference in neuroimaging: theory. Neuroimage 16:465–483.

Hayasaka S, Nichols T (2003): Validating cluster size inference: random field and permutation methods. Neuroimage 20:2343–2356.

Holmes A, Blair R, Watson J, Ford I (1996): Nonparametric analysis of statistic images from functional mapping experiments. J Cereb Blood Flow Metab 16:7–22.

Kherif F, Poline J-B, Mériaux S, Benali H, Flandin G, Brett M (2004): Group analysis in functional neuroimaging: selecting subjects using similarity measures. Neuroimage 20:2197–2208.

Mériaux S, Kherif F, Roche A, Brett M, Garnero L, Poline J-B (2004): How frequently do we sample inhomogeneous group of subjects in fMRI studies? Budapest, Hungary: In: Proc 10th HBM CD-Rom, Neuroimage 22(1).

Neumann J, Lohmann G (2003): Bayesian second-level analysis of functional magnetic resonance images. Neuroimage 20:1346–1355.

Nichols T, Hayasaka S (2003): Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat Methods Med Res 12:419–446.

Nichols T, Holmes A (2002): Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Pallier C, Dehaene S, Poline J-B, LeBihan D, Argenti A, Dupoux E, Mehler J (2003): Brain imaging of language plasticity in adopted adults: can a second language replace the first? Cereb Cortex 13:155–161.

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.

Woolrich M, Behrens T, Beckmann C, Jenkinson M, Smith S (2004): Multi-level linear modelling for fMRI group analysis using Bayesian inference. Neuroimage 21:1732–1747.

Worsley K, Marett S, Leelin P, Vandal A, Friston K, Evans A (1996): An unified statistical approach of determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73.

Worsley K, Liao C, Aston J, Petre V, Duncan G, Morales F, Evans A (2002): A general statistical analysis for fMRI data. Neuroimage 15:1–15.

APPENDIX

EM Algorithm

We detail here the EM algorithm used to maximize the likelihood function given by Eq. (9). The algorithm derives mechanically from the general EM paradigm [Dempster et al., 1977] when considering the collection of true individual effects $(\beta_{1k}, \beta_{2k}, \ldots, \beta_{nk})$ as hidden variables. Given initial estimates $\hat{\beta}_{Gk}$ and $\hat{\sigma}_{Gk}$, the algorithm iteratively refines them by alternating two steps, the E-step (expectation) and the M-step (maximization), until convergence. In our implementation, initial estimates are respectively taken as the classical sample mean and sample standard deviation of the observed effects $(\hat{\beta}_{1k}, \hat{\beta}_{2k}, \ldots, \hat{\beta}_{nk})$.

E-step. Assume current estimates are exact and compute the posterior joint distribution of all subject's true effects. Since subjects are conditionally independent, this reduces to computing each subject's posterior, $p(\beta_{ik}|\hat{\beta}_{ik}, \hat{\sigma}_{ik}, \beta_{Gk}, \sigma_{Gk})$ which is a Gaussian with parameters (m_{ik}, s_{ik}):

$$m_{ik} \leftarrow \frac{\sigma_{Gk}^2}{\hat{\sigma}_{ik}^2 + \hat{\sigma}_{Gk}^2}\,\hat{\beta}_{ik} + \frac{\hat{\sigma}_{ik}^2}{\hat{\sigma}_{ik}^2 + \hat{\sigma}_{Gk}^2}\,\hat{\beta}_{Gk},\; s_{ik} \leftarrow \frac{\hat{\sigma}_{ik}\hat{\sigma}_{Gk}}{\sqrt{\hat{\sigma}_{ik}^2 + \hat{\sigma}_{Gk}^2}} \tag{A.1}$$

M-step. Update $(\beta_{Gk}, \sigma_{Gk})$ by maximizing the expected log-likelihood of the complete data:

$$Q(\beta_{Gk}, \sigma_{Gk}) = n \log \sqrt{2\pi}\sigma_{Gk} + \frac{1}{2\sigma_{Gk}^2} \sum_i [s_{ik}^2 + (\beta_{Gk} - m_{ik})^2], \tag{A.2}$$

yielding:

$$\hat{\beta}_{Gk} \leftarrow \frac{1}{n} \sum_i m_{ik}, \quad \hat{\sigma}_{Gk}^2 \leftarrow \frac{1}{n} \sum_i [s_{ik}^2 + (\hat{\beta}_{Gk} - m_{ik})^2] \tag{A.3}$$

This algorithm is guaranteed to converge towards the (unique) likelihood maximizer $(\hat{\beta}_{Gk}, \hat{\sigma}_{Gk})$over $\mathbb{R} \times \mathbb{R}_+$. In the constrained problem subject to $\beta_{Gk} = 0$, the algorithm is identical except that $\hat{\beta}_{Gk}$ is forced to zero in the M-step instead of being updated. Typically a dozen of iterations are needed to achieve a 1% tolerance on likelihood variations. Faster EM variants probably exist, as discussed by Worsley et al. [2002] in a closely related context.