

Neuroinformatics

IN THIS ISSUE

The Impact of the NIH
Public Access Policy
on Literature Informatics

Statistical Criteria
in fMRI Studies
of Multisensory Integration

Comparison of Vector Space
Model Methodologies to
Reconcile Cross-Species
Neuroanatomical Concepts

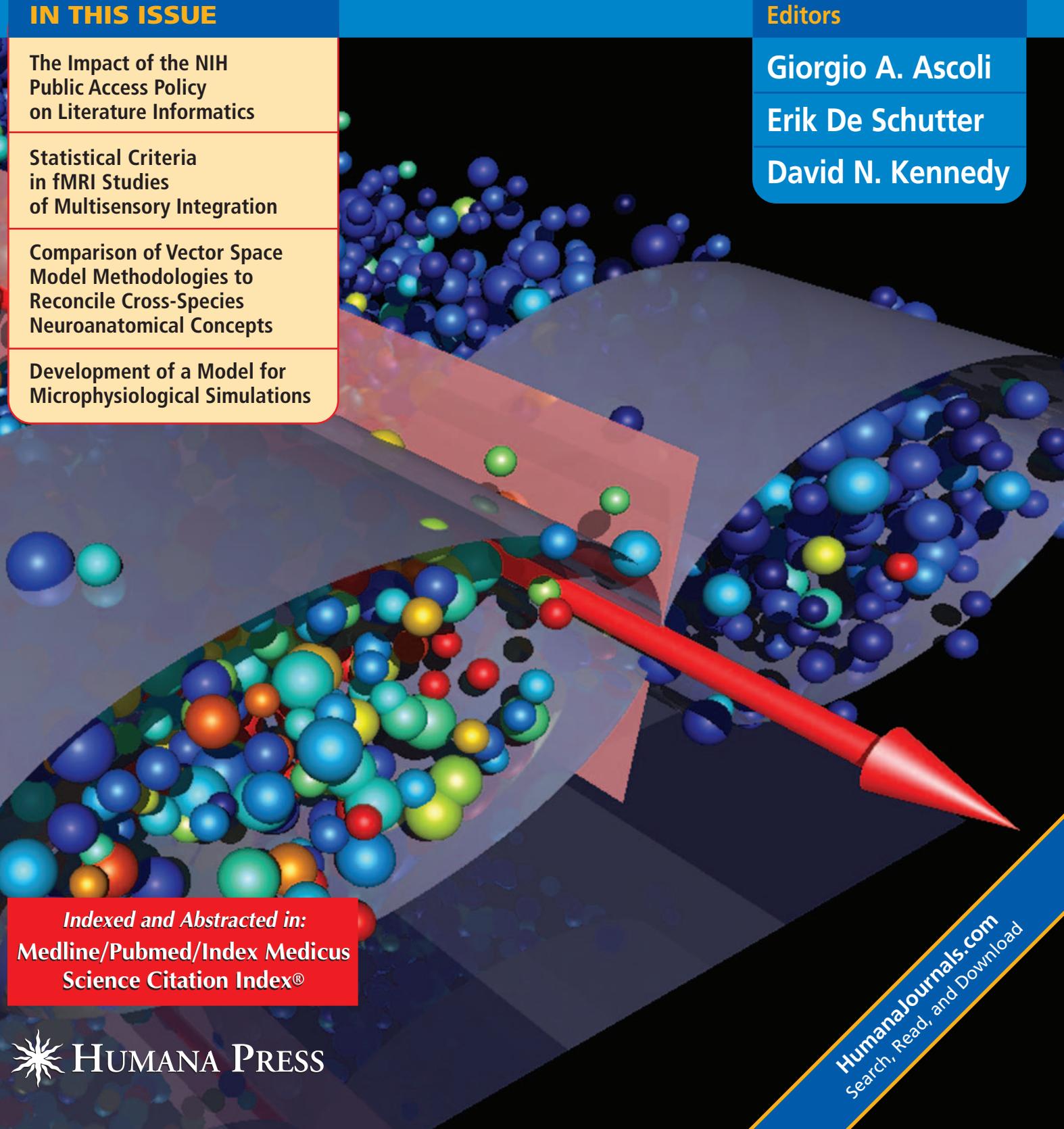
Development of a Model for
Microphysiological Simulations

Editors

Giorgio A. Ascoli

Erik De Schutter

David N. Kennedy



Indexed and Abstracted in:
Medline/Pubmed/Index Medicus
Science Citation Index®

 HUMANA PRESS

Humanajournals.com
Search, Read, and Download

Original Article

Statistical Criteria in fMRI Studies of Multisensory Integration

Michael S. Beauchamp

Laboratory of Brain and Cognition, National Institute of Mental Health Intramural Research Program, National Institutes of Health, Department of Health and Human Services, Bethesda, MD.
E-mail: Michael.S.Beauchamp@uth.tmc.edu

Abstract

Inferences drawn from functional magnetic resonance imaging (fMRI) studies are dependent on the statistical criteria used to define different brain regions as "active" or "inactive" under the experimental manipulation. In fMRI studies of multisensory integration, additional criteria are used to classify a subset of the active brain regions as "multisensory." Because there is no general agreement in the literature on the optimal criteria for performing this classification, we investigated the effects of seven different multisensory statistical criteria on a single test dataset collected as human subjects performed auditory, visual, and auditory-visual object recognition. Activation maps created using the different criteria differed dramatically. The classification of the superior temporal sulcus (STS) was used as a performance measure, because a large body of converging evidence demonstrates that the STS is important for auditory-visual integration.

A commonly proposed criterion, "supra-additivity" or "super-additivity", which requires the multisensory response to be larger than the summed unisensory responses, did not classify STS as multisensory. Alternative criteria, such as requiring the multisensory response to be larger than the maximum or the mean of the unisensory responses, successfully classified STS as multisensory. This practical demonstration strengthens theoretical arguments that the super-additivity is not an appropriate criterion for all studies of multisensory integration. Moreover, the importance of examining evoked fMRI responses, whole brain activation maps, maps from multiple individual subjects, and mixed-effect group maps are discussed in the context of selecting statistical criteria.

DOI: 10.1385/NI:03:02:93

Index Entries: Auditory; visual; multisensory; integration; fMRI; superior temporal sulcus.

Introduction

Functional magnetic resonance imaging (fMRI) studies acquire data from thousands of brain locations, leading to the analysis problem of deciding which brain locations are involved in the experimental task. To solve this problem, statistical analyses are performed on the MR time series. In the most common analysis method, the so-called general linear model*, regressors are created that correspond to the temporal sequence of different events experienced by the subject during scanning (Worsley and Friston, 1995). Brain regions in which the MR time series is time-locked to these events are classified as "active."

A second problem in fMRI data analysis involves assigning the active brain regions to different functional roles. This problem is particularly acute in studies of multisensory integration. In a typical multisensory study in which subjects are presented with stimuli, make cognitive decisions about them, and then produce a motor response, activations would be expected in unisensory regions responding to the sensory stimulus, multisensory regions that integrate across modalities, cognitive regions that are important for decision making, and response selection and motor regions that produce the behavioral output. Even without an explicit behavioral task, subjects may perform language and memory operations when presented with a multisensory stimulus (and without a task there is no measure of subjects' alertness or attention, or the amount of processing performed on each stimulus). With or without a task, multisensory fMRI experiments typically find activity in many brain regions. Multiple statistical criteria are applied to the fMRI data in order to classify a subset of these

active regions as being specifically involved in multisensory integration. A number of criteria have been proposed for this purpose, including the super-additivity requirement (requiring that the multisensory response be larger than the sum of the unisensory responses), the mean requirement (requiring that the multisensory response be larger than the mean of the unisensory responses), and the max requirement (requiring that the multisensory response be larger than the maximum of the unisensory responses) (Calvert, 2001; Calvert et al., 2001).

Because the criteria used to classify active regions as "multisensory" can have a significant impact on the neuroscience inferences that are drawn, it is important to understand the effects of different criteria. It might seem possible to isolate these effects by examining published studies. However, differences in multisensory criteria across studies are confounded by other differences in experimental methods, including stimuli, tasks, subject populations (including anatomical variability across subjects), and imaging hardware. Furthermore, published studies understandably focus on reporting the neuroscience result of interest, with relatively little information about the statistical criteria and processing steps used to obtain the result.

In contrast, the present manuscript is meant as a tutorial to illustrate the effects of individual statistical criteria and processing steps, applied in sequence to a single test dataset. Each step of the analysis is illustrated in detail, while the neuroscience results are reported elsewhere (Beauchamp et al., 2004a,b). As other variables, including the data itself, are held constant, any differences observed are solely because of the analysis method. While data from the entire brain are analyzed, special attention is paid to the superior temporal sulcus (STS),

* In the field of statistics, the general linear model is referred to as the linear model or the multivariate linear model. The acronym "GLM" in statistics has been used for several decades to refer to the generalized linear model (e.g., McCullagh and Nelder, 1983) which is a non-linear method quite different from the linear model.

because subjects in the test dataset performed an auditory–visual object recognition task, and converging anatomical, physiological, and neuroimaging evidence demonstrates that STS plays a crucial role in auditory–visual integration in humans and macaque monkeys (reviewed in Beauchamp, 2005).

This treatment is neither meant to serve as a comprehensive review of the multisensory fMRI literature nor as an exhaustive mathematical analysis of multisensory statistical criteria. Instead, by illustrating the results of applying several commonly used criteria to the same dataset, qualitative understanding of their effects can be gained. This is useful for evaluating published studies (studies A and B seem to have discrepant findings, but viewed from the perspective of their different analysis techniques, they are actually quite compatible). This approach does not allow any definitive conclusions to be drawn about the “correct” statistical criteria, as different brain regions may or may not meet given criteria in any particular experiment, depending on the exact combination of stimulus, task, experimental design, and measurement power.

These illustrations should make it easier for multisensory scientists to perform fMRI experiments without having to perform trial-and-error analysis comparisons on their own data. Intended as a primer, the general guidelines presented for analysis and display of multisensory fMRI data should provide a useful roadmap for choosing statistical criteria and making correct inferences regardless of the exact experimental protocol.

Methods

Human Subjects and MR Data Collection

Eight subjects underwent a complete physical examination and provided informed consent. Subjects were compensated for participation in the study and anatomical MR scans were screened by the NIH Clinical Center,

Department of Radiology in accordance with the NIMH human subjects committee. MR data were collected on a General Electric 3 Tesla scanner. A high-resolution anatomical sequence (one to three repetitions) was collected at the beginning of each scanning session. Gradient-echo echo-planar volumes that were sensitive to blood oxygenation level-dependent (BOLD) contrast were collected following the high-resolution anatomical images. BOLD scan series were acquired with echo time (TE) of 30 ms, repetition time (TR) of 3 s, and 3.75 mm in-plane resolution. Each volume contained 24 axial slices (slice thickness of 4.5 or 5.0 mm as necessary to cover the entire cortex) with 132 volumes per scan series and 8–10 scan series per subject.

Auditory and Visual Stimuli

Subjects viewed visually presented videos of moving manipulable objects (e.g., a hammering hammer), heard recordings of the objects (e.g., ‘bang-bang-bang’) or simultaneously viewed and heard the objects. Visual stimuli were presented using Matlab (Mathworks Inc., Natick, MA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) running on a Macintosh G4 (Apple Computer, Cupertino, CA). Auditory stimuli were presented at approx 80-dB SPL, using a SilentScan system from Avotec, Inc. (Stuart, FL) which attenuates gradient noise produced by the MR scanner while providing high-fidelity stimulus reproduction. Subjects reported being able to hear the stimuli in the scanner and performed the behavioral discrimination task with high accuracy. For additional details, please see Beauchamp et al. (2002, 2003, 2004b).

An event-related design was used. Each trial began with the presentation of a single stimulus (2.5-s duration) followed by a 2.5-s delay, followed by a 3-s display containing three visually presented words. Subjects pressed a button corresponding to the name of the stimulus presented (e.g., hammer/saw/telephone). Each 8-s trial was

separated from the next trial by 0–6 s of fixation baseline. Different trial types were ordered for optimal experimental efficiency (Dale, 1999) using the optseq program written by Doug Greve. In this ordering, the stimuli were presented at varying times on a 1-s time base, allowing estimation of the hemodynamic response to a single stimulus of each type with 1-s resolution.

fMRI Data Analysis

MR data were analyzed within the framework of the linear model in AFNI 2.50 (Cox, 1996). The first two volumes in each scan series, collected before equilibrium magnetization was reached, were discarded. Then, all volumes were registered to the third volume in the first scan series, because it was collected nearest in time to the high-resolution anatomy. Next, a spatial filter with a root-mean-square width of 4 mm was applied to each echo-planar volume. All volumes from each subject were concatenated to form a single extended time series.

In different brain regions, the response to each trial contained BOLD responses to either the sensory stimulus of (beginning at $t = 0$ s) the motor response (occurring at approx $t = 6$ s) or both (Fig. 2). Because each trial contained two separate events, the common assumption that the BOLD response should take the form of a single γ -variate or gaussian function was inappropriate. Instead, a method known as deconvolution, also called finite impulse response (FIR), was used to analyze the response. A separate regressor was used to model the response in each 1-s period in a 20-s window following trial onset. With three stimulus types, this resulted in 60 regressors of interest, each consisting of a series of δ functions (equivalent to boxcars with a width of 1 s [11], the resolution of the estimation procedure) that were then fit to the MR time series separately for each voxel and each subject. Each regressor independently represented 1 s of the time-locked evoked response, hence no shape constraints were placed on the calculated

response, and it was free to form a two-peaked distribution (or any other shape). The analysis produced a 20-s evoked response for each trial type in each voxel for each subject. While this evoked response is assumption-free (besides the usual FIR requirements of linearity and time invariance) it must be related to the underlying neural activity. By assuming that the hemodynamic BOLD signal peaks 4–6 s after neural activity, the amplitude of the neural response to the stimulus (presented from $t = 0$ to 3 s) can be estimated by summing the beta-weights (also known as the fit parameters or amplitudes) of the regressors representing the fifth to the eighth second of the response. In effect, this calculates the area under the curve (summed total) of the largest part of the evoked BOLD hemodynamic response and uses it as a measure of the sum of the underlying neural activity. The amplitude of the neural signal related to the motor response (approx $t = 6$ s) was estimated by summing the 11th to the 14th second of the BOLD response.

The regression analysis calculated an overall experimental effect that described how much variance in each voxel time series was accounted for by all regressor of interest (the 60 δ functions). This experimental effect F -ratio was thresholded at $p < 10^{-6}$ (to correct for the multiple comparisons produced by approx 20,000 brain voxels) in order to create individual subject activation maps that showed all voxels with a time-locked response to any trial component. Following stringent thresholding by this F -ratio, voxels were categorized by their response to different trial components using a more liberal threshold of $p < 0.05$, described below. Functional data were interpolated to 1 mm³ resolution using cubic interpolation and overlaid on single subject anatomical data.

To create group maps, a mixed-effects model was used. For each subject, the regression model provided a single estimate of the response to each stimulus type in each voxel. After stereotactic normalization to Talairach space (Talairach

and Tournoux, 1988), a two-way ANOVA was performed on each voxel in standard space.

The same contrasts were used for both single subject and group analysis. In single subjects, the amplitude of the response to each type of trial was calculated by summing the amplitudes of three regressors: fifth to eighth seconds of auditory trials formed the auditory response, fifth to eighth seconds of visual trials formed the visual response, and fifth to eighth seconds of multisensory trials formed the multisensory response. In order to compare the different trial types, contrasts were performed between the amplitudes of these values. If the trial types are represented Auditory Visual Multisensory, super-additivity was measured with the contrast $-1 - 11$ while sub-additivity was measured with the contrast $1 1 - 1$. The mean criterion was instantiated with the contrast $-1 - 1 2$. The max criterion was generated by selecting the minimum t -value from two separate contrasts of M vs A $-1 0 1$ and M vs V $0 -1 1$.

Surface Modeling

Three-dimensional models of the cortical surfaces were constructed using FreeSurfer software. One to five high-resolution anatomical scans for each subject were collected and averaged. An automated segmentation routine then extracted the gray-white boundary and constructed a surface model, which was then inflated to allow the inspection of active areas buried deep in cortical sulci (Fischl et al., 1999). Surfaces were visualized using SUMA software (Argall et al., in press).

Results

Single Subject Activation Maps

Even at a conservative statistical threshold of $p < 10^{-6}$, a widely distributed set of brain regions in frontal, parietal, occipital, and temporal lobes showed a significant modulation of the MR time series time-locked to trial onsets (shown for a single representative subject in Fig. 1A). Trials

contained separate sensory stimulation and behavioral task epochs (Fig. 2A–C) allowing independent estimation of the brain activity during auditory stimulation (A), visual stimulation (V), auditory-visual multisensory stimulation (M), and motor response (R) epochs. By examining the evoked MR response from different regions of interest (Fig. 2) we can assess the sensitivity of different brain regions to the different trial types and epochs.

Initial Criterion: Remove Global Deactivations

One subset of active regions in medial frontal and medial parietal cortex showed a decreased MR signal (below fixation baseline) in all experimental conditions (Fig. 2D). These same set of regions have been found in many neuroimaging studies to be active ‘at rest’, that is, during any task with little sensory input or motor output, such as fixation baseline. This has led to the suggestion that these regions may be a resting state network, important for internal reflections on past or future events that are suppressed during any type of goal-directed behavior (Raichle et al., 2001). Because fMRI measures the BOLD signal change between conditions, the amplitude (and sign) of these deactivations is dependent on the control and experimental conditions used. For instance, a control task that is cognitively demanding will eliminate the deactivations, but has the disadvantage of activating an entirely new set of regions, muddying the comparison with the experimental condition of interest. Therefore, the most common approach in multisensory imaging studies is to use a well-studied control task (such as a passive fixation baseline). This allows the set of deactivated regions to be predicted accurately.

Because these regions are deactivated in most sensory tasks, they are not thought to be specifically involved in sensory processing. Their inclusion can confuse further analysis.

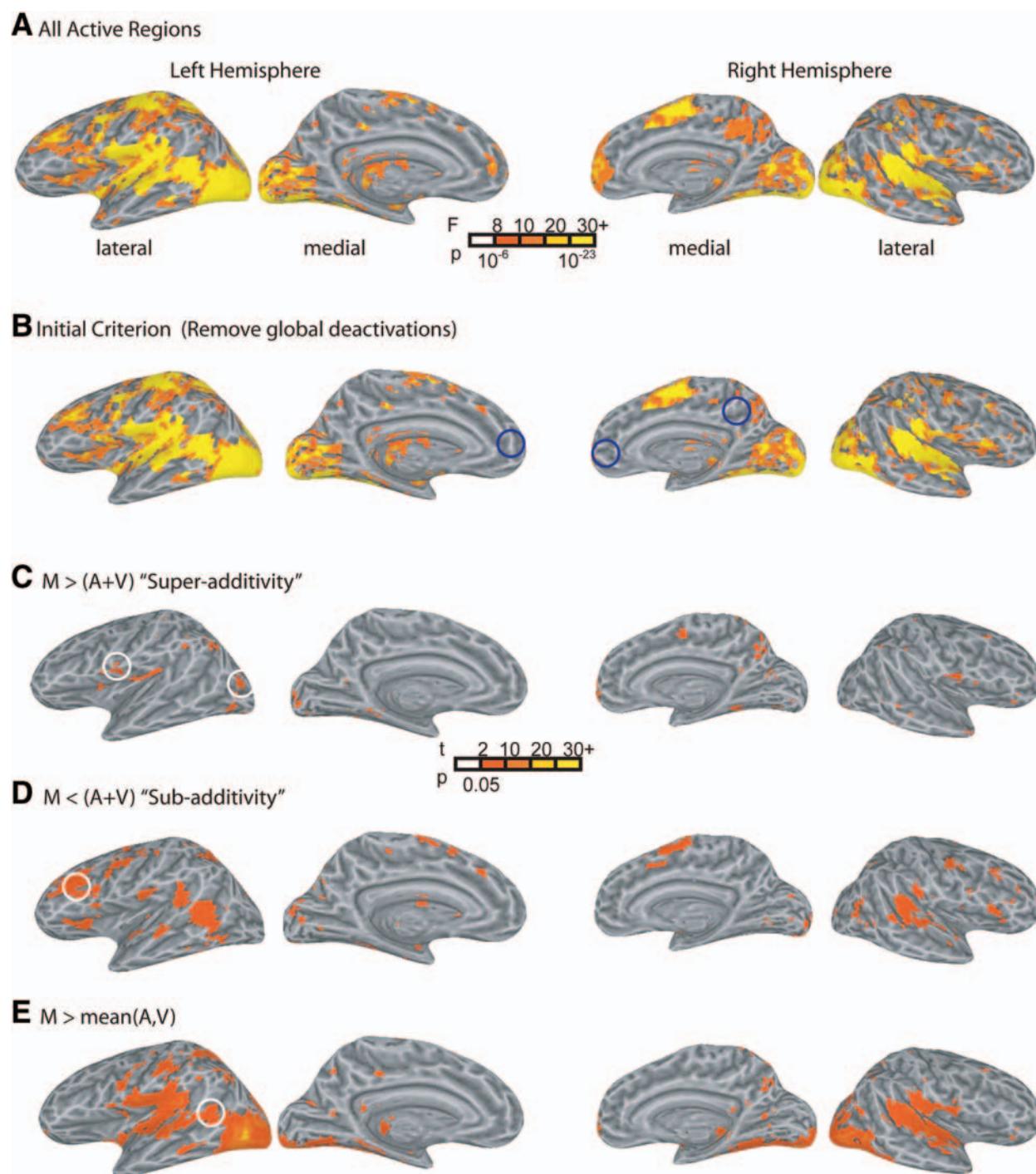


Fig. 1. Activation maps from a single subject using different multisensory statistical criteria. Lateral and medial views of the inflated left and right hemisphere are shown. Colors indicate functional data (all colored region pass statistical criteria) mapped to the cortical surface. Grayscale indicates anatomical structure with dark gray corresponding to sulcal depths, light gray to gyral crowns. **(A)** All regions showing a significant experimental

For instance, if a map is made of the contrast between visual and auditory response magnitudes, the deactivated regions would be shown as visually preferring because the deactivation is smaller during visual than auditory epochs (Fig. 2D, $V = -0.5\%$ vs $A = -0.75\%$). This might give rise to the erroneous conclusion that the deactivated regions are part of a visual processing network. Therefore, deactivated regions are often removed before further analysis. In equation form, a criterion to do this is $(A > 0) \text{ OR } (V > 0) \text{ OR } (M > 0) \text{ OR } (R > 0)$. Applying this criterion to each voxel independently produces an activation map with fewer active regions, all of which show a positive signal change during some portion of the evoked response (Fig. 1B; blue circles illustrate regions of the resting state network removed by this criterion). It is important to note that this is not a blanket prescription to ignore all deactivations. For instance, different categories of stimuli can modulate the amplitude of deactivations (Mitchell et al., 2002) meaning that it is critical to carefully examine the responses to all multisensory conditions, even in deactivated regions (Laurienti et al., 2002).

Next, we examine the remaining active regions that are classified as multisensory by additional criteria (Fig. 1C–E). The first additional criterion, super-additivity (Fig. 1C) requires that the multisensory response be

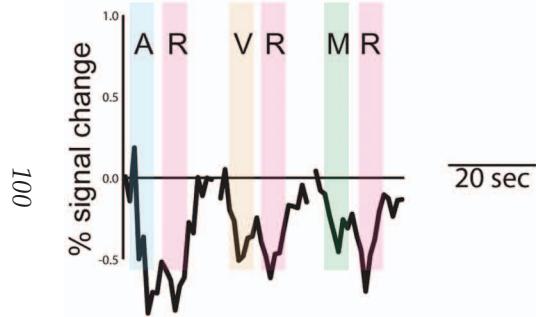
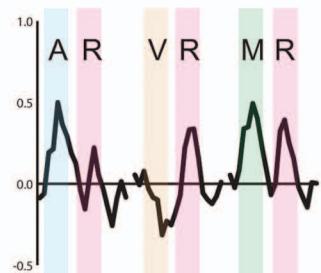
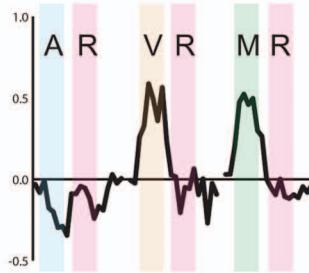
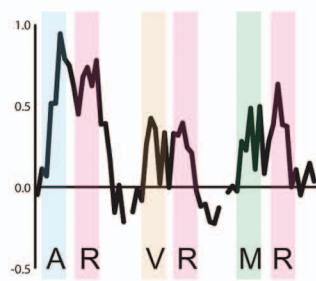
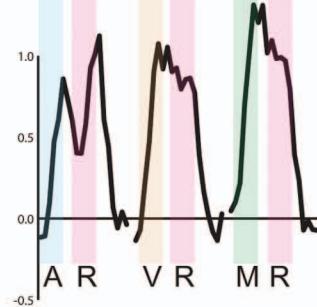
greater than the sum of the individual unisensory responses. We can represent this in equation form as $M > (A + V)$. Note that this is applied in serial with the initial criterion, so the complete criterion is $((A > 0) \text{ OR } (V > 0) \text{ OR } (M > 0) \text{ OR } (R > 0)) \text{ AND } (M > (A + V))$. Because we wish to attach the statistical significance to the contrast of $M > A + V$, we visualize only brain regions that show a significant ($t > 2$, $p < 0.05$) value for this contrast. As illustrated in Fig. 1C, this criterion produces a broadly scattered pattern of relatively small activations. Notably, no activation is observed in the STS. If we examine the MR time series from regions in auditory and visual association cortex that do show a super-additivity effect, we find that this effect largely arises from deactivations during the auditory and visual conditions. For instance, auditory cortex (Fig. 2E) shows an auditory response of $A = 0.5\%$, a visual response of $V = -0.2\%$, for a predicted multisensory response of $(0.5\%) + (-0.2\%) = 0.3\%$. Because the true multisensory response is $M = 0.5\%$, the criterion of $M > (A + V)$ is satisfied. In visual association cortex (Fig. 2F), a similar effect is observed ($A = -0.2\%$, $V = 0.6\%$, $M = 0.5\%$).

Because most active regions did not fulfill the super-additivity criterion of $M > (A + V)$, many regions show a significant effect of the reverse contrast, $M < (A + V)$, or sub-additivity (Fig. 1D). Examining the time series from one

effect. Color bar illustrates significance of experimental effect for (A) and (B). **(B)** All regions showing a significant experimental effect and a positive change from fixation baseline during any experimental condition. Blue circles indicate regions removed by this criterion (evoked MR response from these regions shown in Fig. 2D). **(C)** Regions from (B) with the additional criterion that the multisensory auditory–visual response (M) must be greater than the sum of the auditory (A) and visual (V) responses. Circles illustrate activity in auditory and visual association cortex (evoked MR responses from these regions shown in Fig. 2E,F). Color bar shows significance of M vs A + V contrast. **(D)** Regions from (B) with the criterion that the multisensory response is less than the summed auditory and visual responses. Circle illustrates activity in dorsolateral pre-frontal cortex (evoked MR response shown in Fig. 2G). Colors show significance of A + V vs M contrast. **(E)** Regions from (B) with the criterion that the multisensory response is larger than the mean of the auditory and visual response. Circle illustrates activity in superior temporal sulcus (STS) (evoked MR response shown in Fig. 2H). Colors show significance of M vs mean(A,V) contrast.

A Auditory Trial**B Visual Trial****C Multimodal Trial**

Average MR time series from regions of interest

D Medial frontal/parietal**E Auditory assoc. ctx****F Visual assoc. ctx****G DLPFC****H STS**

of these regions, in dorsolateral prefrontal cortex (Fig. 2G), we find a large activation in the auditory condition with smaller activations in the visual and auditory–visual condition, so that $M < A + V$.

For the third additional criterion (Fig. 1E), we examine the difference between multisensory and unisensory responses with the contrast $M > \text{mean}(A + V)$. This contrast detects multisensory activation in STS and adjacent unisensory auditory and visual cortex. The STS response takes the form of positive auditory and visual responses, with a larger response during multisensory stimulation (Fig. 2H).

Initial Criterion: Remove Global Deactivations and Unisensory Deactivations

As shown in the time series from auditory and visual association cortex (Fig. 2E,F), the MR signal in early sensory areas decreased below fixation baseline during stimulation in the nonpreferred modality. Because increased neural activity is usually taken as the hallmark of information processing, multisensory fMRI studies often eliminate areas showing deactivation from consideration by requiring that regions show a positive BOLD response during both unisensory auditory and visual stim-

ulation. We may represent this initial criterion as $(A > 0) \text{ AND } (V > 0)$. This criterion removes a number of active brain regions (Fig. 3A). Like the global deactivation criterion, it eliminates the resting state network (because for these areas, $A < 0$ and $V < 0$). Unlike the global deactivation criterion, this criterion also eliminates motor cortex and other brain regions involved only in the behavioral response, because for these regions $R > 0$ but approx $A = 0$ and $V = 0$. Note that this effect is unique to the event-related design of the test dataset in which sensory stimulus (A or V) is separated in time from behavioral response (R). In most studies, stimulus and response occur in the same experimental epoch (so that A is actually $A + R$ and 'V' is actually $V + R$) meaning that motor cortex and related regions would pass the criterion of $A > 0$ and $V > 0$.

The largest region of activation that passed the unisensory deactivation initial criterion was located in left posterior STS (arrows in Fig. 3A) along with other smaller regions of activation in frontal and parietal cortex. As before, we apply additional criteria in series with this initial criterion. No brain regions passed the super-additivity criterion (Fig. 3B). In contrast, the mean criterion classified the large STS region as multisensory (Fig. 3C). A subset of this region passed the max criterion.

Fig. 2. Details of trial structure and evoked MR responses. **(A)** Each auditory trial consisted of a 2.5-s auditory stimulus (A , blue bar) followed by a 2.5-s delay followed by a 3-s response period (R , purple bar). **(B)** Visual trial consisted of a visual stimulus (yellow bar, V) followed by delay and response periods. **(C)** Multisensory trials consisted of a simultaneous auditory and visual stimulus (green bar, M) followed by delay and response periods. **(D)** Mean evoked MR response from medial frontal and parietal voxels exhibiting signal decreases below fixation baseline in every condition. Location of voxels are shown with blue circles in Fig. 1B. Responses to each trial type are shown sequentially, although they were presented in pseudorandom order separated by varying intertrial intervals of fixation baseline (not shown). **(E)** Evoked response from voxels in auditory association cortex exhibiting super-additivity (left circle in Fig. 1C). **(F)** Evoked response from voxels in visual association cortex exhibiting super-additivity (right circle in Fig. 1C). **(G)** Evoked response from voxels in dorsolateral prefrontal cortex (DLPFC) exhibiting sub-additivity (Fig. 1D). **(H)** Evoked responses from STS voxels exceeding the mean criterion M (Fig. 1E).

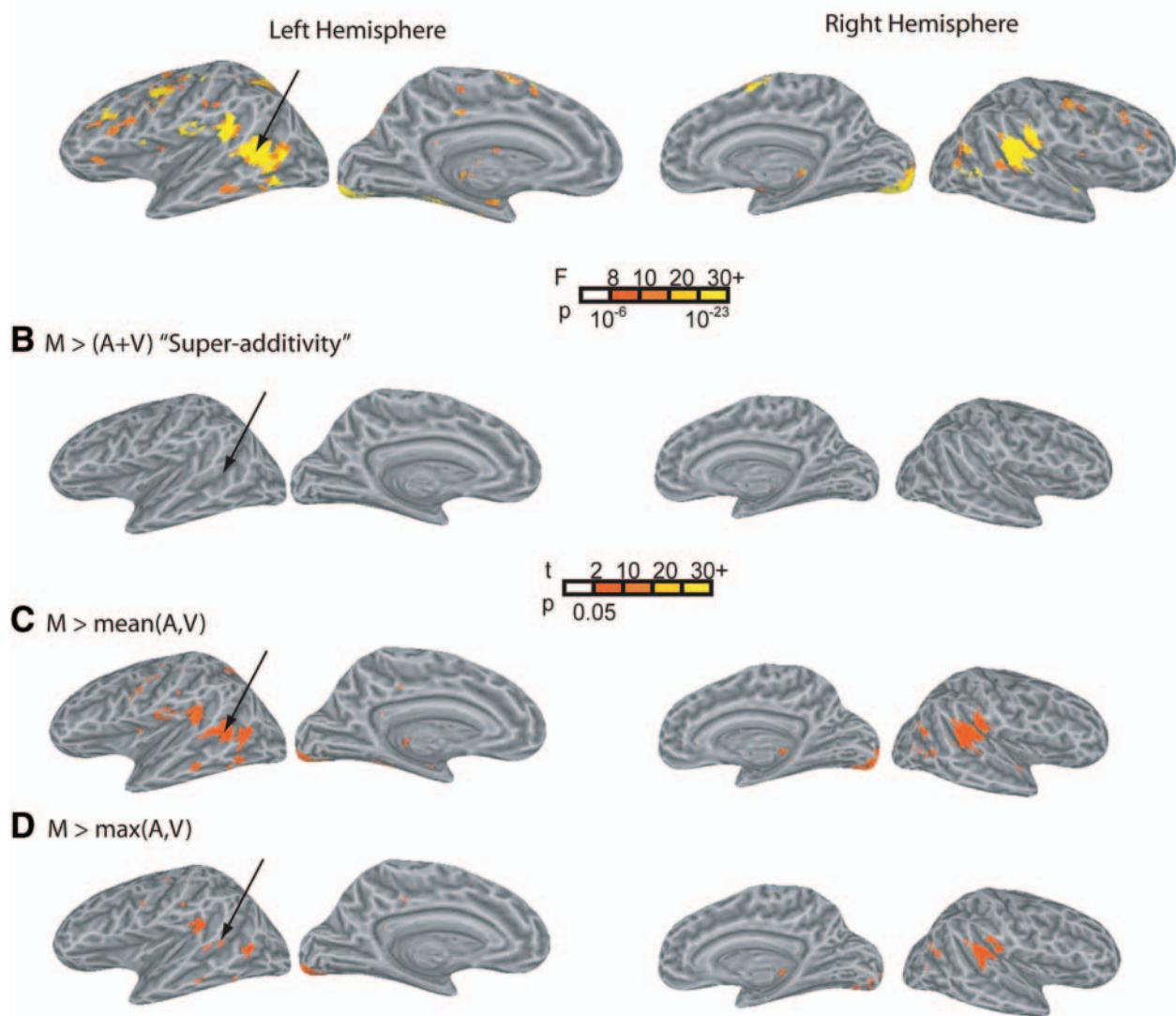
A Initial Criterion (Remove global and unimodal deactivations)

Fig. 3. Activation maps from a single subject using different multisensory statistical criteria. Display conventions as in Fig. 1. **(A)** All regions showing a significant experimental effect (Fig. 1A) and a positive change from fixation baseline during auditory or visual stimulation epochs ($A > 0$ OR $V > 0$). Black arrow indicates large focus of activity in STS. Color bar indicates significance of experimental effect. **(B)** Regions from (A) with the additional criterion of super-additivity. No voxels pass the criterion. **(C)** Voxels from (A) that also exceed the mean criterion. Color bar show significance of M vs $\text{mean}(A,V)$ contrast. **(D)** Voxels from (A) exceeding the max criterion. Colors show significance of M vs $\text{max}(A,V)$ contrast.

Group Activation Maps

On an individual subject basis, applying different criteria to a multisensory dataset pro-

duced very different activation maps. Because intersubject variability in fMRI is high, the effects of different statistical criteria on group average activation maps were examined.

Initial Criterion: Remove Global Deactivations

Many brain regions were active in the average activation map (Fig. 4A). The initial criterion removed resting state regions in medial frontal and parietal cortex (Fig. 4B). Examining the results of the additional criteria, only a few scattered voxels were super-additive (Fig. 4C) while many brain regions passed the inverse criterion of sub-additivity (Fig. 4D). The mean criterion identified regions in STS and nearby unisensory auditory and visual cortex (Fig. 4E).

Initial Criterion: Remove Global Deactivations and Unisensory Deactivations

As in the single subject case, the initial criterion was modified to remove unisensory deactivations (Fig. 5A). When combined with this criterion, super-additivity failed to identify any active voxels (Fig. 5B). In contrast, a large region of STS was classified as multisensory by the mean criterion. Only a few scattered voxels were identified as multisensory by the max criterion.

Intersubject Variability and the Effect of Threshold

Figure 6A illustrates activation maps from four normal subjects performing the same auditory–visual object identification task with similar behavioral performance. Brain regions that passed the unisensory deactivation and mean criteria were identified in each subject. The general outline of activity was similar across subjects, with multisensory activity focused in the posterior STS. However, there was also significant intersubject variability.

Even for a given subject and statistical criterion, the exact threshold chosen also has an impact on the resulting activation map (Fig. 6B,C). By increasing the statistical threshold of the initial criterion (Fig. 6B), we can produce successively smaller regions of activation (com-

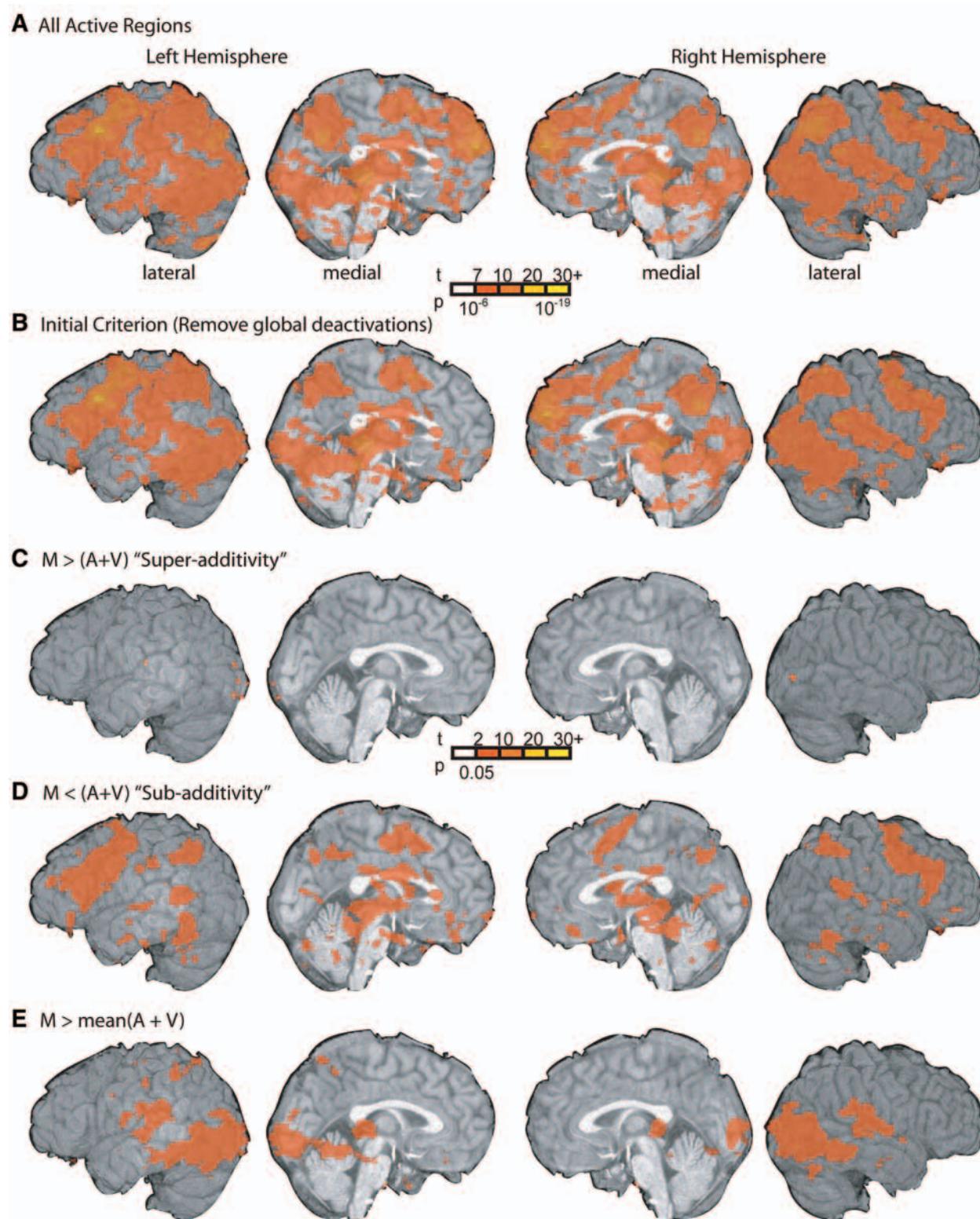
pare with Fig. 3A). By increasing the statistical threshold of the secondary criterion (in this example the max criterion), the number of regions meeting criterion is successively reduced (Fig. 6C).

Importance of Multiple Thresholding Steps

Both the individual subject and group analyses that were discussed earlier contain two discrete thresholding steps (following the removal of deactivated regions). In the first step, all active brain regions are selected by thresholding the overall significance of the regression model (omnibus *F*-test) at a stringent significance level ($p < 10^{-6}$). In the second step, the multisensory statistical criterion (e.g., AV vs A + V) is applied at a much lower significance ($p < 0.05$). The first step amounts to a Bonferroni correction for multiple comparisons (because the brain volume contains 10,000–100,000 voxels, each of which is tested separately, many false positives will result if a high threshold is not used). Because the second criterion is applied only to the small pool of voxels that passed the first threshold, a correction for multiple comparisons is not applied.

This two-step process is widely used in fMRI studies. For instance, in a study examining face and house selective regions in ventral temporal visual cortex (Haxby et al., 1999), the first thresholding step was used to detect all voxels that were visually responsive at a high threshold ($z > 4, p < 6 \times 10^{-5}$) while the second step detected voxels that preferred faces or houses at a much lower threshold ($z > 1.96, p < 0.05$).

The fundamental rationale for the two-step method can be understood if we consider that the difference in percentage signal change between conditions (e.g., AV vs A + V; faces vs houses) is relatively small compared to the difference between all conditions and fixation baseline. In the two-step method, the large difference between experimental conditions and



baseline is used for the stringent contrast that corrects for multiple comparisons while the more liberal threshold is used to detect the subtle difference between experimental conditions.

An obvious alternative to this approach is to use one-step thresholding process on only the comparison of interest, such as (AV vs A + V) or (faces vs houses). Figure 6D illustrates the difficulties inherent in this approach. If a significance level appropriate to correct for multiple comparisons ($p < 10^{-6}$) is selected for the one-step threshold, very few voxels exceed the significance threshold because of the relatively small difference in percent signal change between experimental conditions. However, if a more liberal threshold is selected for the one-step analysis many false positive regions are detected (for instance, outside the brain or in white matter). Therefore, additional analysis must be performed to attempt to separate true activations from false positive activations. One common approach is to adjust the significance level by the size of the observed activations (Xiong et al., 1995). For example, in a random dataset, the probability of two adjacent voxels exceeding a significance of $p < 0.05$ is $0.05^2 = 0.0025$. In practice, there is considerable spatial correlation in an fMRI dataset. This means that adjacent voxels are not truly independent, and may form activation clusters that consist entirely of false positive voxels (e.g., a cluster of correlated voxels in white matter). In addition, adjusting thresholds in this man-

ner imposes *a priori* assumptions about the expected size of the activation. This can be problematic if the structure being examined, such as the superior colliculus in multisensory studies, is anatomically restricted and therefore only occupies a small number of voxels. The two-step threshold process avoids these difficulties because it does not require assumptions about correlation between voxels or the size of active regions.

Discussion

fMRI studies of multisensory integration usually examine one (or a few) brain regions. These regions are selected from a larger set of active regions using criteria that classify some brain regions as multisensory and discard others. Dramatic differences were observed in the brain regions classified as multisensory depending on the statistical criteria used.

Because the different criteria have similar degrees of face validity, particular care was taken to examine the results of the classification process in STS. There is overwhelming evidence, based on a large body of anatomical and physiological evidence from both human and nonhuman primates (reviewed in Beauchamp, 2005), that STS is a key cortical area for auditory-visual integration. If this fact is taken as an *a priori* assumption (to avoid circular logic) criteria that successfully classify STS as multisensory are more likely to be useful.

Fig. 4. The mixed-effects group average map from eight subjects created using different multisensory statistical criteria. Volume renderings of lateral and medial views of left and right hemisphere are shown. Colors indicate average functional data, grayscale indicates anatomical structure from a single subject. Activations with $|x| > 30$ are visualized in the lateral rendering, activations with $|x| < 30$ are visualized in the medial rendering. **(A)** All regions showing a significant experimental effect. Color bar illustrates significance of experimental effect for (A) and (B). **(B)** All regions showing a significant experimental effect and a positive change from fixation baseline during any experimental condition. **(C)** Regions from (B) with the additional criterion of super-additivity. Color bar shows significance of M vs A + V contrast. **(D)** Regions from (B) with the additional criterion of sub-additivity. Color bar shows significance of A + V vs M contrast. **(E)** Regions from (B) that also exceed the mean criterion. Color bar show significance of M vs mean(A,V) contrast.

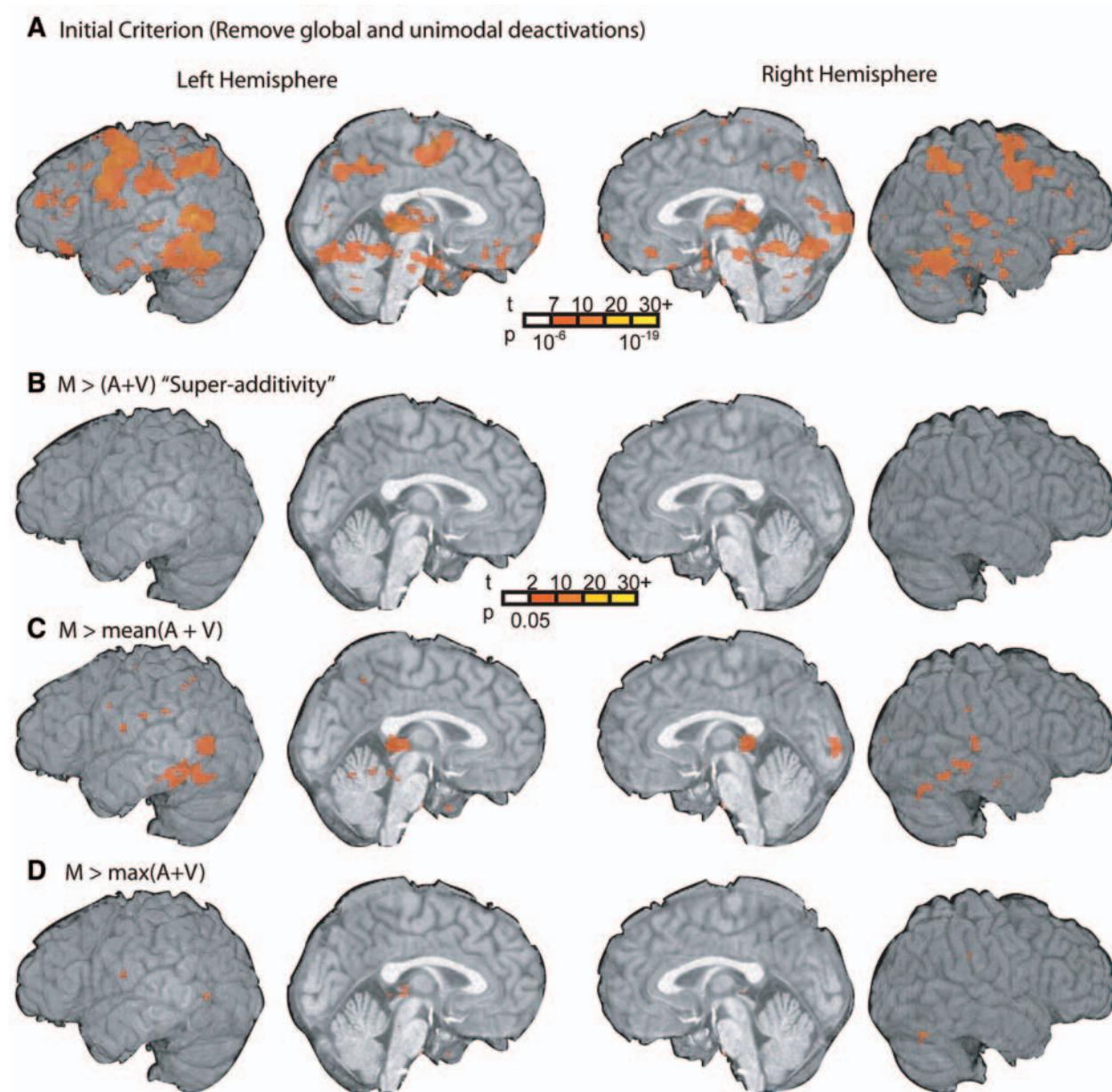


Fig. 5. The mixed-effects group average map created using different multisensory statistical criteria. Display conventions as in Fig. 4. **(A)** All regions showing a significant experimental effect (Fig. 4A) and a positive change from fixation baseline during auditory or visual stimulation epochs ($A > 0$ OR $V > 0$). Color bar indicates significance of experimental effect. **(B)** Regions from (A) with the additional criterion of super-additivity. No voxels pass the criterion. **(C)** Voxels from (A) that also exceed the mean criterion. Color bar show significance of M vs $\text{mean}(A,V)$ contrast. **(D)** Voxels from (A) exceeding the max criterion. Colors show significance of M vs $\text{max}(A,V)$ contrast.

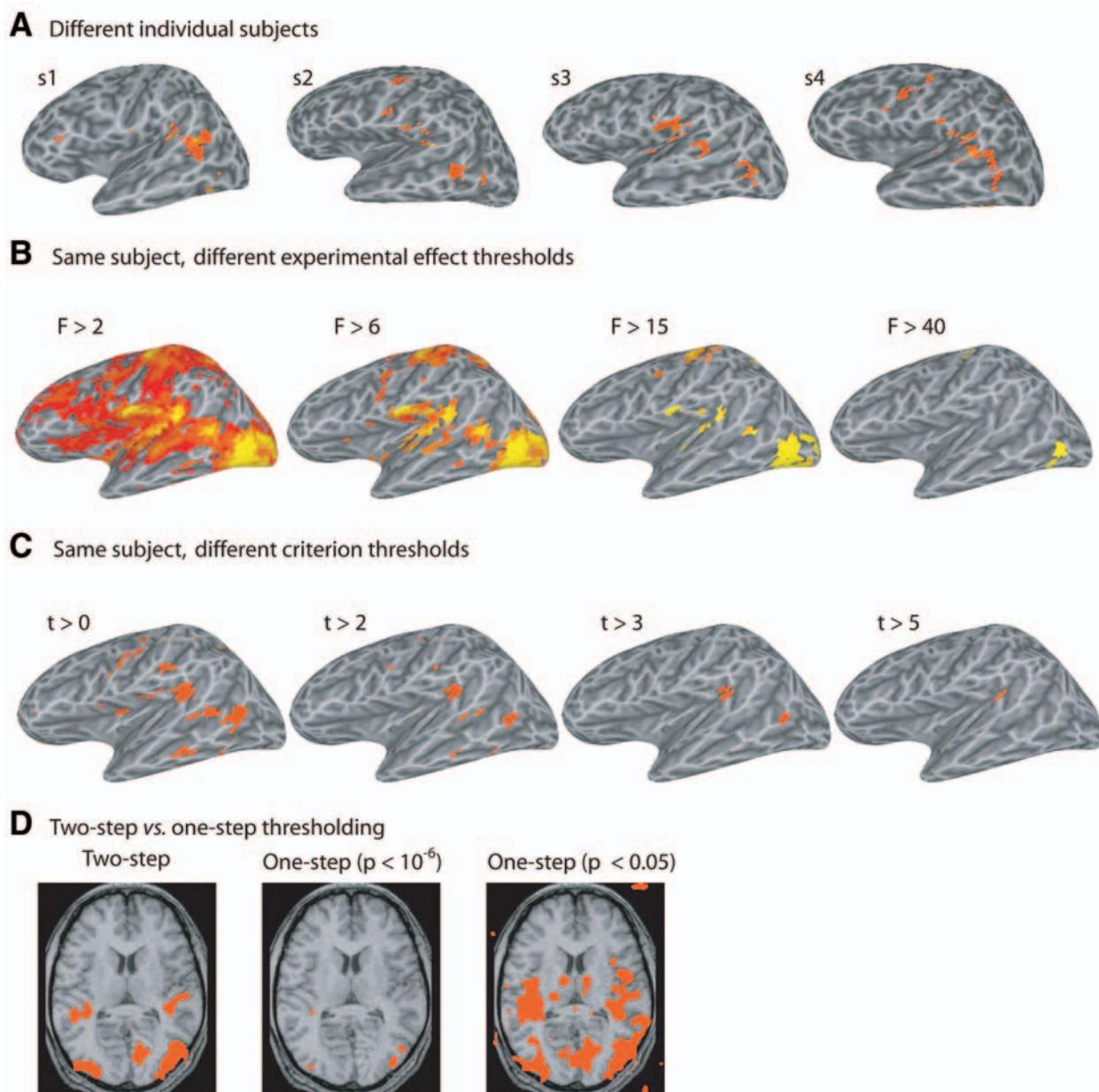


Fig. 6. Single subject activation maps from different subjects, with different statistical thresholds, and with one or two statistical criteria. **(A)** Lateral views of the left hemisphere from four subjects illustrating regions showing a significant experimental effect, a positive change from fixation baseline during auditory or visual stimulation epochs, and meeting the mean criterion $M > \text{mean}(A, V)$. Compare with Fig. 3C for a fifth subject. **(B)** Lateral view of a single subject left hemisphere at four different thresholds for the overall experimental effect, measured as an F -ratio. **(C)** Lateral view of a single subject left hemisphere at four different thresholds for the t -statistic of the contrast M vs $\text{max}(A, V)$. Overall experimental effect threshold set to $F > 8$. **(D)** Axial slice ($z = 10$) through an individual subject using a two-step analysis, a one-step analysis with high threshold, or a one-step analysis with liberal threshold.

A Perspective on Statistical Criteria

When we speak of a brain area performing a task, such as multisensory integration, it is tempting to imagine a single cytoarchitecturally defined region with a relatively small size. It is certainly possible to create maps with a single tiny focus by constructing activation maps using extremely high thresholds (Fig. 6). However, most evidence suggests that multiple brain areas participate nearly in all cognitive events, including processing simple sensory stimuli. While some areas show smaller signal changes than others, with repeated averaging even areas of weak activation can pass stringent statistical tests (Saad et al., 2003).

This can be understood if we imagine the activation profile in each active region as a mountain with a tall peak and a broad base. At high statistical thresholds, or in data with low signal-to-noise ratio (SNR), only the very tip of the peak will reach significance, resulting in extremely focal activation patterns. At lower thresholds (or with repeated averaging to give high SNR) much more of the mountain will exceed significance, resulting in larger active regions (Fig. 6B).

Statistical criteria for multisensory integration can be considered in this light. The super-additivity criterion can be written as $M > A + V$ and the mean criterion can be written as $M > (A + V)/2$. Therefore, both criteria are operating on the same comparison (M vs $A + V$) but with different degrees of strictness. Because super-additivity is more strict, it inevitably leads to smaller regions of multisensory activation, sometimes eliminating them completely (Fig. 3B).

Advantages and Disadvantages of Different Secondary Criteria

At a theoretical level, super-additivity is attractive because it proposes using the same criterion for fMRI studies ($M > A + V$) that has been applied in recording studies of multisensory neurons (Calvert et al., 2001). However, as emphasized by Laurienti and colleagues (Laurienti et al., 2005) while some individual multisensory neurons are super-additive, others are sub-additive ($M < A + V$). In BOLD fMRI, the combined response of many neurons in a voxel is measured. Because super- and sub-additive neurons exist in similar numbers, and are not spatially segregated, the single neuron data do not make clear predictions about the expected BOLD response to multisensory stimulation. Therefore, the super-additivity criterion may be overly strict, and introduce type II (false negative) error. Evidence for this was found in the present analysis, in which multisensory activity in STS was not detected with super-additivity. The converse criterion of sub-additivity was also proposed on the basis of single-neuron studies and is subject to the same concerns about the correspondence between single-neuron and BOLD responses. As observed in the present analysis, sub-additivity suffers from the drawback of being relatively nonspecific. For instance, a voxel in motor cortex that shows the same percent signal change in each condition, e.g., $(M, A, V) = (1\%, 1\%, 1\%)$ will be classified as sub-additive because $M < (A + V)$.

If super-additivity is used as a criterion, it is important that the evoked responses be closely examined. For instance, Wright and colleagues (Wright et al., 2003) examined responses to audiovisual speech, and found that some regions that were classified as over-additive (their term for super-additivity) showed substantial deactivations during visual stimulation. Different routes to super-additivity, e.g., $(M, A, V) = (3\%, 1\%, 1\%)$ vs $(M, A, V) = (1\%, 1\%, -1\%)$ likely reflect different underlying neural processes.

The maximum criterion $M > \max(A, V)$ can also be written $(M > A) \text{ AND } (M > V)$ (Van Atteveldt et al., 2004). Qualitatively, it is less strict than super-additivity and more strict than the mean response criterion (compare Fig. 3C–E). One disadvantage of the max criterion is that it is

nonlinear, so that small perturbations in the amplitude of either unisensory response can give large differences in the calculated criterion. For example, the weaker of the two unisensory responses plays no role in determining the max criterion, so that two voxels with very different response patterns, such as $(M,A,V) = (5\%,4\%,4\%)$ and $(M,A,V) = (5\%,4\%,1\%)$ are assigned the same max criterion although their responses are quite different.

Unlike the max criterion, the mean criterion is linear and well behaved because it reflects the contribution of both unisensory responses. The contrast of M vs $\text{mean}(A,V)$ provides a useful index of the degree of multisensory integration in an area. In some respects, it tests the null hypothesis that the response across conditions is similar. For instance, a voxel in motor cortex with response $(M,A,V) = (1\%,1\%,1\%)$ will show a small value for the mean criterion, whereas a voxel with $(M,A,V) = (2\%,1\%,1\%)$ will show a large value. Because it is more liberal than the super-additivity criteria, it is able to identify multisensory regions, including STS. A disadvantage of the mean criterion is that it may be too liberal. Without an initial criterion requiring unisensory activation in each modality (i.e. $A > 0$ AND $V > 0$), the mean criterion classifies unisensory areas as multisensory, e.g., if $(M,A,V) = (1\%,0\%,1\%)$, $\text{mean}(A + V) = 0.5\% < 1\%$.

Advantages and Disadvantages of Different Initial Criteria

The initial unisensory deactivation criterion of $(A > 0 \text{ AND } V > 0)$ has the advantage of removing early sensory areas that are not traditionally thought to be multisensory. However, recent research shows that deactivations in early sensory cortex may reflect early multisensory processing (Laurienti et al., 2002). Even areas that are unisensory by definition, such as primary visual cortex, may receive multisensory input (Falchier et al., 2002; Foxe et al., 2002; Schroeder et al., 2003). In most studies, it is reasonable to weed out regions that are deactivated

to one or all sensory stimuli. However, given the uncertainties surrounding these issues, it is important to perform at least a cursory examination of all areas without any initial criteria.

Reasons for Careful Examination of fMRI Data

Regardless of the criteria used, one must carefully examine activation maps and MR time series at each processing stage. This is vital because fMRI analysis software is complex, and its operations and algorithms are incompletely understood by most users. This is true even for well-documented, widely used software, such as SPM (Worsley and Friston, 1995). Consider the example of conjunction analysis (Friston et al., 1999). Conjunction analysis is commonly used in multisensory studies to find brain regions that are active under every experimental condition, that is, respond to stimuli in every sensory modality (e.g., Bremmer et al., 2001). However, as emphasized by Nichols et al. (2005), conjunction analysis in SPM99 and SPM2 actually tests the hypothesis that a brain region is active in *no* experimental condition. Hence, a positive result for a region in a conjunction analysis of auditory, visual, and tactile tasks does *not* indicate that the region was responsive to all three modalities. A good way to detect errors of this type is to examine the evoked MR response in each region in each condition.

Summary and General Guidelines for Analysis and Display of Multisensory fMRI Data

In the test dataset used for the present comparison, the super-additivity criterion failed to detect multisensory activity in STS, despite the use of a high-field scanner and a sophisticated event-related design. Multisensory integration is enhanced when the unisensory stimuli are ambiguous (e.g., Alais and Burr, 2004). In the test dataset, subjects were relatively accurate in both the unisensory auditory (79%, chance 33%) and unisensory visual

(92%) conditions. While subjects were more accurate in the multisensory condition (94%), this represents a small improvement, perhaps corresponding to the small degree of multisensory enhancement observed in the fMRI signal. Other well-powered studies that examined multisensory integration using unambiguous unisensory stimuli have also failed to detect super-additivity in STS (Amedi et al., 2001, 2002; Van Atteveldt et al., 2004). These results strengthen the theoretical arguments discussed earlier (Laurienti et al., 2005), suggesting that super-additivity is not always an appropriate criterion in multisensory fMRI studies. Super-additivity could be effective when used in an experimental design in which the demands of the stimulus and task result in a high degree of multisensory enhancement. In general, the proper criteria may depend on the exact details of the experiment being analyzed and the neuroscience hypothesis being tested. Regardless of the exact analysis method used, several useful guidelines can help to ensure that correct inferences are drawn from multisensory fMRI data.

Average MR Time Series

Examining the evoked MR response from regions of interest (plotted as % signal change vs time, Fig. 2) across different trial types allows verification that the statistical criteria applied have performed as expected. At a more basic level, it allows verification that the evoked responses have the shape typical of BOLD hemodynamic responses (relatively slow changes that peak 4–6 s after the expected neural activity). The additional information available in the time dimension of the evoked response permits an estimation of the SNR ratio of the experiment, which is not possible with simple bar charts or scatter plots of activation intensities. Variance in the evoked response can be illustrated with error bars at each time-point demonstrating the

standard error or standard deviation across subjects (e.g., Beauchamp et al., 2003).

Whole Brain Activation Maps

Early fMRI papers often illustrated activity with a single slice through the brain volume. This was sometimes owing to hardware limitations (the ability to collect data from only a single slice) or the desire to only show activity in a single region, obviating the need to discuss activity in other brain regions. With technological advances, modern fMRI studies almost always collect data from a large fraction of the brain. Because these data are collected and analyzed, it is appropriate to illustrate all active regions that meet the analysis criteria, regardless of their location in the volume. This can be accomplished with a variety of techniques: cortical surface modeling (Fig. 1) volume rendering (Fig. 4) or simply multiple slices through the volume (e.g., Petit and Beauchamp, 2003).

Mixed-Effects Group Maps

In a mixed-effects model, stimulus is the fixed factor (all subjects view the same stimuli) while individual subject is the random factor (meaning that the degrees of freedom is equal to the number of subjects). The mixed-effects method is sometimes called random-effects, to distinguish it from the less-favored fixed-effects approach in which subjects are not treated as a factor (meaning that the degrees of freedom is artificially high, equal to the total number of brain volumes collected).

Activation Maps from Multiple Individual Subjects

As shown in Fig. 6, active areas can differ significantly in number and location from subject to subject. The mixed-effects group average activation map captures commonalities in the activation pattern across subject, but does not completely describe the activation. For instance, a relatively focal patch of activation in the group average activation

map may reflect a consistent focal region of activation in individual subjects. Or, it may reflect the result of averaging a number of subjects with quite different activation patterns (in terms of number of active foci and size of activations) that overlap maximally at the location of the peak focus in the average map. For a full understanding of the brain architecture underlying a cognitive task, it is necessary to examine several individual subject activation maps in addition to the group average map.

Behavioral Data

Some form of behavioral task is an important element of multisensory fMRI studies. Attention has dramatic effects on the BOLD response (e.g., Beauchamp et al., 1997) and BOLD responses are weaker during passive presentation of stimuli than during active performance of a behavioral task with the same stimuli (e.g., Beauchamp et al., 1999). More subtly, differences in task difficulty can cause changes in brain activation that can be falsely attributed to other task differences. For instance, auditory object recognition tasks are often more difficult than visual object recognition tasks. Since greater task difficulty is associated with greater activation in frontal regions, these regions would be more active during auditory object recognition and could be wrongly classified as auditory-preferring cortex. However, with behavioral data showing longer reaction time or lower percent correct during auditory trials, this confound could be noted and corrected for (perhaps by making the auditory task easier or the visual task more difficult).

Comprehensive Description of Analysis Methods

Even within prepackaged analysis software, there is a combinatorially vast number of ways to process fMRI data. A clear and complete description of the methods is vital. The recent

widespread adoption of supplementary online materials makes an even more comprehensive description possible, including depiction of the results of various alternative analysis strategies.

Information Sharing Statement

The software described in this manuscript are available as follows: stimulus presentation software from <http://lbc.nimh.nih.gov/people/mikeb/matlab.html>; AFNI and SUMA from <http://afni.nimh.nih.gov/afni/download>; FreeSurfer from <http://surfer.nmr.mgh.harvard.edu/download.html>; the optseq program from <http://surfer.nmr.mgh.harvard.edu/optseq/>. The data and stimuli can be obtained via personal communication with the author.

Acknowledgments

The author wishes to thank the organizers of the fifth annual meeting of the International Multisensory Research Forum, especially Dr Salvador Soto-Faraco, for providing the opportunity to engage in many discussions (including with Drs Paul Laurienti and Tommi Raij) that gave intellectual spur to this manuscript. This research was supported by the National Institute of Mental Health Intramural Research Program.

References

- Alais, D. and Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262.
- Amedi, A., Jacobson, G., Hendler, T., Malach, R., and Zohary, E. (2002) Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cereb. Cortex* 12, 1202–1212.
- Amedi, A., Malach, R., Hendler, T., Peled, S., and Zohary, E. (2001) Visuo-haptic object-related activation in the ventral visual pathway. *Nat. Neurosci.* 4: 324–330.
- Argall, B. D., Saad, Z. S., and Beauchamp, M. S. A simplified method for intersubject averaging on the cortical surface using SUMA. *Human Brain Mapping*, in press.

- Beauchamp, M. S. (2005) See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr. opin. Neurobiol.* 15, 145–153.
- Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., and Martin, A. (2004a) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.* 7, 1190–1192.
- Beauchamp, M. S., Cox, R. W., and DeYoe, E. A. (1997) Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *J. Neurophysiol.* 77, 516–520.
- Beauchamp, M. S., Haxby, J. V., Jennings, J. E., and DeYoe, E. A. (1999) An fMRI version of the Farnsworth–Munsell 100-Hue test reveals multiple color-selective areas in human ventral occipitotemporal cortex. *Cereb. Cortex* 9, 257–263.
- Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004b) Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41, 809–823.
- Beauchamp, M. S., Lee, K. E., Haxby, J. V., and Martin, A. (2002) Parallel visual motion processing streams for manipulable objects and human movements. *Neuron* 34, 149–159.
- Beauchamp, M. S., Lee, K. E., Haxby, J. V., and Martin, A. (2003) fMRI responses to video and point-light displays of moving humans and manipulable objects. *J. Cogn. Neurosci.* 15, 991–1001.
- Brainard, D. H. (1997) The psychophysics toolbox. *Spat. Vis.* 10, 433–436.
- Bremmer, F., Schlack, A., Shah, N. J., et al. (2001) Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. *Neuron* 29, 287–296.
- Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123.
- Calvert, G. A., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001) Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage* 14, 427–438.
- Cox, R. W. (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173.
- Dale, A. M. (1999) Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* 8, 109–114.
- Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002) Anatomical evidence of multimodal integration in primate striate cortex. *J. Neurosci.* 22, 5749–5759.
- Fischl, B., Sereno, M. I., and Dale, A. M. (1999) Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *NeuroImage* 9, 195–207.
- Foxe, J. J., Wylie, G. R., Martinez, A., et al. (2002) Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *J. Neurophysiol.* 88, 540–543.
- Friston, K. J., Holmes, A. P., Price, C. J., Buchel, C., and Worsley, K. J. (1999) Multisubject fMRI studies and conjunction analyses. *NeuroImage* 10, 385–396.
- Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., and Martin, A. (1999) The effect of face inversion on activity in human neural systems for face and object perception. *Neuron* 22, 189–199.
- Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y. F., Field, A. S., and Stein, B. E. (2002) Deactivation of sensory-specific cortex by cross-modal stimuli. *J. Cogn. Neurosci.* 14, 420–429.
- Laurienti, P. J., Perrault Jr., T. J., Stanford, T. R., Wallace, M. T., and Stein, B. E. (2005) On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp. Brain Res.*, in press.
- McCullagh, P. & Nelder, J. A. (1983). Generalized Linear Models, Vol. 37 of Monographs on Statistics and Applied Probability, 1st edition, Chapman and Hall, London.
- Mitchell, J. P., Heatherton, T. F., and Macrae, C. N. (2002) Distinct neural systems subserve person and object knowledge. *Proc. Natl. Acad. Sci. USA* 99, 15238–15243.
- Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J. B. (2005) Valid conjunction inference with the minimum statistic. *NeuroImage* 25, 653–660.
- Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442.
- Petit, L. and Beauchamp, M. S. (2003) Neural basis of visually guided head movements studied with fMRI. *J. Neurophysiol.* 89, 2516–2527.
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001) A default mode of brain function. *Proc. Natl. Acad. Sci. USA* 98, 676–682.
- Saad, Z. S., Ropella, K. M., DeYoe, E. A., and Bandettini, P. A. (2003) The spatial extent of the BOLD response. *NeuroImage* 19, 132–144.
- Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N., and Hackett, T. A. (2003)

- Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *Int. J. Psychophysiol.* 50, 5–17.
- Talairach, J. and Tournoux, P. (1988) Co-planar stereotaxic atlas of the human brain, Thieme Medical Publishers, New York.
- Van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004) Integration of letters and speech sounds in the human brain. *Neuron* 43, 271–282.
- Worsley, K. J., and Friston, K. J. (1995) Analysis of fMRI time-series revisited - again. *NeuroImage* 2, 173–181.
- Wright, T. M., Pelpfrey, K. A., Allison, T., McKeown, M. J., and McCarthy, G. (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex* 13, 1034–1043.
- Xiong, J. H., Gao, J. H., Lancaster, J. L., and Fox, P. T. (1995) Clustered pixels analysis for functional MRI activation studies of the human brain. *Hum. Brain Mapp.* 3, 287–301.

