Published in final edited form as:
*Brain Lang*. 2007 August ; 102(2): 176–185. doi:10.1016/j.bandl.2006.04.015.

# Test-retest reliability in fMRI of language: Group and task effects

**E. Elinor Chen**[a,*] and **Steven L. Small**[a,b,c]

a*Department of Radiology, Brain Research Imaging Center, The University of Chicago, USA*

b*Department of Neurology, Brain Research Imaging Center, The University of Chicago, USA*

c*Department of Psychology, Brain Research Imaging Center, The University of Chicago, USA*

## Abstract

This paper explores how the test-retest reliability is modulated by different groups of participants and experimental tasks. A group of 12 healthy participants and a group of nine stroke patients performed the same language imaging experiment twice, test and retest, on different days. The experiment consists of four conditions, one audio condition and three audiovisual conditions in which the hands are either resting, gesturing, or performing self-adaptive movements. Imaging data were analyzed using multiple linear regression and the results were further used to generate receiver operating characteristic (ROC) curves for each condition for each individual subject. By using area under the curve as a comparison index, we found that stroke patients have less reliability across time than healthy participants, and that when the participants gesture during speech, their imaging data are more reliable than when they are performing hand movements that are not speech-associated. Furthermore, inter-subject variability is less in the gesture task than in any of the other three conditions for healthy participants, but not for stroke patients.

### Keywords

Language; Brain; BOLD; Brain imaging; Test-retest reliability; Receiver operating characteristic (ROC) curves; fMRI; Neurological disease

## 1. Introduction

Since the advent of functional magnetic resonance imaging (fMRI) about 10 years ago (Ogawa et al., 1992), this method has become a primary tool for elaborating brain/behavior relationships in a variety of sensory/motor and cognitive domains and in a variety of normal and impaired populations. Typically, an investigator will make inferences from a collection of brain images, where the nature of such images depends crucially on the statistical analysis performed. If a statistical result exceeds a (more or less arbitrary) threshold, it will be concluded there is experimentally induced activation in an individual voxel or collection of voxels (i.e., a brain region).

Due to the noninvasive character of fMRI, it is becoming a useful tool for studying patients with neurological illnesses. For example, fMRI is playing a role in pre-operative planning (Fernández et al., 2003) as well as in the evaluation of neurological recovery after stroke (Binkofski & Seitz, 2004; Cao, D'Olhaberriague, Vikingstad, Levine, & Welch, 1998; Cramer et al., 1997; Pineiro, Pendlebury, Johansen-Berg, & Matthews, 2002; Small, Hlustik, Noll,



* Corresponding author. Fax: +1 773 834 7610. *E-mail address:* E-mail: chene@uchicago.edu (E.E. Chen)..

Genovese, & Solodkin, 2002) and/or during therapeutic intervention (Peck et al., 2004; Small, Flores, & Noll, 1998). The great majority of these studies are based on a single experiment, and it is commonly assumed that the result is reliable, where reliability is defined as the extent to which an activation map is reproducible after repeated experiments. It is important to assess the validity of this assumption, to characterize the factors that affect reliability, and to understand how they do so.

Another reason to study the reliability of fMRI results is the complicated nature of the fMRI signal itself. It is known that the signal comes from the blood oxygen level-dependent (BOLD) contrast effect, and that it indirectly reflects neural activity in the brain. One would expect that different tasks would induce different amounts of neural activity, that the brains (or regions of brains) of different people would manifest different hemodynamic properties on performing the same task, or even that the brain of a single person would behave differently when performing the same task at two different times. Consequently, it would be expected that the BOLD response could be task, participant, and time dependent.

The magnitude of the typical BOLD response is about 1-5%, which is not that different from the magnitude of the baseline noise of fMRI time series. Such noise has multiple sources, elaborated by several authors (for a review, see Genovese, Noll, & Eddy, 1997), but is generally thought to include physiological noise (e.g., cardiac and respiratory-related signal changes), field drifts and other equipment instability, subject movement, environmental effects (e.g., vibration), and others. Unfortunately, due to the complexity of signal and multi-source noise, their overall statistical properties are (mostly) unknown. Consequently, different statistical processing methods can also affect reproducibility of fMRI results.

Many studies have been conducted to investigate how these diverse factors affect reliability of the fMRI signal. Certain studies have focused on the day-to-day reproducibility of an experimental result for a single participant, concluding that such variability ought not to be neglected (McGonigle et al., 2000; Noll et al., 1997; Wei et al., 2004). Further investigations have suggested that the location of activation is reproduced more reliably than are activation magnitude and extent (Marshall et al., 2002; Waldvogel, Gelderen, Immisch, Pfeiffer, & Hallett, 2000). When different subjects are asked to perform the same task, the experimental result is more easily reproducible within any one subject than across multiple subjects (Aguirre, Zarahn, & D'Esposito, 1998). The majority of these studies explore reliability on either a single task (Cohen & DuBois, 1999; Wei et al., 2004; Xiong et al., 2000) or multiple tasks from different domains, such as visual and motor tasks (McGonigle et al., 2000; Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000; Waldvogel et al., 2000). There are only a few studies (Liou et al., 2003; Rutten, Ramsey, Rijen, & Veelen, 2002; Specht, Willmes, Shah, & Jaäncke, 2003) that investigate similar tasks from within a single domain, and very few studies (Manoach et al., 2001) comparing reliability between healthy participants and patients.

The very notion of reproducibility needs definition and elaboration, particularly given the presence of a number of diverse perspectives. Qualitatively, reproducibility can be characterized by the number of experiments that have the same result out of the total number of the experiments performed. In fMRI, such reproducibility requires establishment of an a priori statistical threshold, with reproducibility defined with respect to that particular threshold. Changing the threshold would change the reproducibility. Furthermore, in fMRI, statistics are performed on a voxel-by-voxel basis, such that reproducibility over two executions of an experiment can be defined by the number of voxels that are always inactive, active on the first execution (the “test”) only, active on the second execution (the “re-test”) only, or active on both executions. Again, this “test-retest” reliability depends on a particular threshold, and can be extended beyond two replications of an experiment to an arbitrary number.

Test-retest reliability can be quantified in a number of ways. For example, *t*-values (Fernández et al., 2003; Specht et al., 2003) of the activated voxels can be used to calculate correlation coeffcients between test and retest for an individual subject. In this case, a correlation coeffcient close to one indicates a good reliability. For a given region, correlation coefficients can also be calculated across individual subjects, thus providing a reliability measure on a group level (Manoach et al., 2001; Miezin et al., 2000). Although these approaches are quite straightforward, they have the disadvantage that the analysis is conducted only among voxels that have survived an a priori thresholding procedure. An alternative approach is to use receiver operating characteristic (ROC) curves, which use data at multiple statistical thresholds. In this way, all voxels in the brain are utilized in the analysis.

Receiver operating characteristic curves originated from signal detection theory, a field of investigation developed to address questions related to the ability of an observer to detect a signal in the presence of noise. There are four possible relationships between the observation and the actual signal: true positive (detecting a signal when it exists), false positive (detecting a signal when it does not exist), true negative (detecting no signal when indeed there is no signal) and false negative (detecting no signal when such a signal does exist). By varying different criteria, for example, loosening the threshold, an observer may make more true positive observations while at the same time also making more false positive decisions. To reflect this trade-off, ROC curves are generated by plotting the fraction of true positives against the fraction of false positives.

As a method to measure detection ability and evaluate observers' performance, ROC analysis has an extensive history in both psychophysics (Green & Swets, 1976) and more recently in image processing (Metz, 1978). Conventional ROC analysis, as used in image processing, depends on having an a priori knowledge of truth, i.e., a “gold standard,” by which it is possible to judge whether a measurement is accurate or inaccurate. In the most characteristic case, an imaging diagnosis is compared to the pathological analysis of a tissue specimen, with the tissue diagnosis providing to radiology the “truth” by which the imaging procedure could aspire. Since there is no way to know the true status, active or inactive, of a single voxel in a functional imaging experiment, fMRI does not have such a gold standard or an independent assessment of truth.

One of the first papers to use ROC analysis to compare different acquisition sequences in fMRI (Constable, Skudlarski, & Gore, 1995) is based on simulation results. Le and Hu (1997) used ROC analysis to compare different analysis strategies by scanning one subject twice, and assuming that the truly active voxels could be determined from the first experiment by using an extremely high threshold. Later Nandy and Cordes (2004) extended this approach by using an additional resting state to ascertain the truly inactive voxels. The main improvement of these approaches to ROC analysis over that of Constable et al. (1995) is that they use real data instead of depending completely on simulation. Nevertheless, the true status of voxels completely depends on a single measurement, which is based on a strong assumption that the result is reproducible. Furthermore, as pointed out in their discussion (Nandy & Cordes, 2004), the validity of the method for high-level cognitive tasks remains unknown.

The present study addresses the following questions: 1. Does reliability of an auditory language task change with the addition of reinforcing visual information? 2. Does the answer to this question depend on the neurological status of the participants, or do people with previous strokes have a different degree of reliability on exactly the same tasks as healthy adults? To address these questions, we studied two groups of participants, with and without previous strokes, and performed the same imaging experiment twice, test and retest, on different days. Test-retest reliability was characterized using the particular ROC analytic method proposed by Genovese et al. (1997). Since event-related experiment design studies have demonstrated

that hemodynamic responses can have different shapes across subjects and even for the same subject at different times (Aguirre et al., 1998), a block designed experiment was used to focus on study of test-retest reliability of magnitude, location, and extent of BOLD response.

## 2. Methods

### 2.1. Subjects and experiment

Twelve healthy subjects (6 females; 21 ± 5 years old) and nine stroke patients (4 females; 19 ± 7 years old) participated in the study. All participants are native English speakers. The stroke patients all had perinatal or early postnatal injuries and normal language, and thus did not have injury to functional language areas. All 12 healthy subjects and four stroke patients were right-handed, and five stroke patients were left-handed, as determined by Edinburgh Handedness Inventory (Oldfield, 1971). All participants participated in two test sessions, test and retest, with a time interval between tests varying from 1 day to almost a year. Details of the time interval between test sessions are given in Table 1 for both healthy participants and stroke patients.

The experimental stimuli consisted of modified versions of Aesop's Fables, with each story lasting between 40 and 45 s in duration. There were four conditions in the experiment. In a speech-only condition participants were instructed to listen to the stories through headphones while viewing a fixation cross projected onto a screen at the foot of the scanner bed. The other three conditions were audio-visual conditions, in which the storyteller was visible on the screen. In a no-gesture condition, the storyteller kept her arms and hands at rest while speaking. The other two conditions were accompanied by arm and hand movements. In a gesture condition, the storyteller was allowed to move her hands naturally during speech. Such speech-associated gestures are a normal accompaniment of speech and play a role in language comprehension (McNeill, 1992). Finally, a self-adaptor condition contained hand movements not typically associated with speech, but common in everyday life, such as adjusting clothes, jewelry, and hair. These four conditions were presented once each in random order in each of two runs. Thus, there were four stories per run, with a total of two runs for each test and four runs total for test and retest.

Brain images were collected on a 3 T GE Signa scanner (GE Medical Systems, Milwaukee, WI) with a standard quadrature GE head coil. Thirty functional images, with slice thickness of 5 mm, were acquired along the sagittal direction using a gradient-echo spiral pulse sequence (Glover & Law, 2001; Noll, Cohen, Meyer, & Schneider, 1995) to cover the whole brain (TR = 2 s; TE = 25 ms; flip angle = 77°; 3.75 mm × 3.75 mm in-plane resolution). A high-resolution structural T1-weighted spoiled gradient-recalled (SPGR) volume was obtained for each individual subject along the sagittal direction at TR/TE of 8 s/3.2 ms with voxel size 0.938 × 0.938 × 1.5 $mm^3$. The length of the individual stories varied slightly, such that the stimulus file for functional acquisition in the first test session consisted of three different versions that were chosen randomly for the subject before scanning: The first run of test data was either 4:40 or 4:46 in duration (i.e., containing 140 or 143 TR's) and the second run was 4:56 in duration (148 TR's). For the second (retest) session, functional data acquisition was of the same duration for all subjects, with the first run taking 4:24 (132 TR's) and the second 4:38 (139 TR's). The length of each story was 48 ± 6 s (26 ± 3 TR's) and the interstimulus interval was 14 or 16 s (7 or 8 TR's), which comprised the control baseline (rest) condition for purposes of analysis.

### 2.2. Image processing

The spiral images were reconstructed off-line and then analyzed as time series of image volumes (Cox, 1996). To correct for motion, the functional images for each subject were registered to the first image using a six-parameter, rigid-body transformation (Cox &

Jesmanowicz, 1999). Following that, functional images were registered to the anatomical images. The same procedure was performed for the retest images. Two steps were required to register the test and retest brain images for an individual subject—first, the structural images from the second test session (retest) were registered to the structural images for the first session (test), and then the rotation matrix used for this registration was applied to the retest functional images. Finally, all registered functional images were smoothed with 5 mm FWHM Gaussian kernel.

Regression analysis was performed on these smoothed images on a voxel-by-voxel basis for each run with a priori estimate of the hemodynamic response. The analysis included three regressors corresponding to the constant, linear, and quadratic terms of the baseline, and another four regressors corresponding to the four conditions of the experiment. To exclude false classification of active voxels due to subject's head movement, a similar regression analysis was also conducted with the addition of motion correction parameters as regressors. The statistical results (*F* value) for each condition were further used to generate reliability maps and estimate false positive fractions and true positive fractions in the ROC analysis.

### 2.3. ROC analysis

For the ROC analysis, we used the mixture binomial model proposed by Genovese et al. (1997). In this model, each voxel in the brain is assumed to be drawn from an independent identical distribution. The probability of a truly active voxel being considered as active (TP) is denoted as $P_A$ and probability of a truly inactive voxel being considered as active (FP) is denoted as $P_I$. The probability that a voxel is classified as active in $j$ out of $M$ replications is given by

$$\binom{M}{j}\left[\lambda P_A^j(1-P_A)^{M-j}+(1-\lambda)P_I^j(1-P_I)^{M-j}\right], \tag{2.1}$$

where λ is the proportion of truly active voxels in the brain. Eq. (2.1) is valid when applied to a single threshold case and has to be modified when multiple thresholds are included in the model. The reason for this is that a voxel considered active at a high threshold is automatically considered active at a lower threshold. Due to the dependence of thresholds, a dependent model was suggested as an extension of above model

$$\prod_{\vec{t}}\left[\lambda\prod_{k=0}^{K}P_{Ak}^{t_k}+(1-\lambda)\prod_{k=0}^{K}P_{Ik}^{t_k}\right]^{n_{\vec{t}}}. \tag{2.2}$$

A constant coefficient is neglected in Eq. (2.2). Index $k$ denotes an increasing threshold level and varies between 0 and $K$ where $K = 7$ in our analysis. These eight thresholds were determined based on F-statistics from the multiple linear regression, and were arranged in increasing order, $F_0 < F_1 < \ldots < F_7$. $P_{Ak}$ ($P_{Ik}$) is the probability that a truly active (inactive) voxel is classified to be active at threshold level $k$, i.e., between thresholds $F_k$ and $F_{k+1}$. Since in all cases, a voxel is assigned to one of the $k$ levels, $P_{Ak}$ and $P_{Ik}$ are subjected to the following constraints: $\sum_{k=0}^{K}P_{Ak}=1$ and $\sum_{k=0}^{K}P_{Ik}=1$. $n_{\vec{t}}$ is a vector, $(t_0, \ldots, t_K)$, whose elements $t_k$ denote the number of times a voxel is classified as active at threshold level $k$. Sum of all $t_k$ equals total number of replications, $M$. After this classification is performed for each single voxel, voxels in the whole brain are then grouped based on classification pattern. For example, $n_{\vec{t}}$ labels the number of voxels, which have the same combination $\vec{t}$. Finally, a maximum likelihood estimation was

used to estimate parameter λ and $P_{Ak}$ and $P_{Ik}$ at each threshold level $k$. For a detailed derivation of Eq. (2.2), please refer to Appendix of Genovese et al. (1997). All of above procedures were coded in MATLAB.

Under the assumption that signal and noise are drawn from bi-normal distributions, parameters $P_{Ak}$ and $P_{Ik}$ were further used as input for a Fortran program (LABROC4) to generate fitted ROC curves. This program was also used to estimate the area under the curve, $A_z$. The fitted ROC curves and estimated parameter $A_z$ were collected for each individual subject under each task condition and were then used for further ROC curve comparisons.

## 3. Results

### 3.1. Qualitative results

Reliability maps were produced by counting the number of times, from the total of four runs, a voxel was classified as active at a given threshold. Fig. 1 illustrates the reproducibility maps of one participant performing the gesture task at three different thresholds. Active voxels are colored blue, yellow, orange, and red, depending on whether they were active in one, two, three, or all four of the runs, respectively. From bottom to top in Fig. 1, the number of red colored voxels, which were considered as active in all four runs, increased as the threshold decreased (or $p$-value increased). This corresponds to intuition: As the threshold is lowered, truly active voxels have a higher probability of being considered active, i.e., the number of true positives is increased. On the other hand, truly inactive voxels also have a greater probability of being considered active, i.e., false positives are also increased. This trade-off between false positives and true positives will become evident in the ROC analysis discussed later.

Fig. 2 shows the reproducibility maps at a given threshold for healthy participants (Fig. 2a) and stroke patients (Fig. 2b) performing all four tasks. The reproducibility maps are quite similar across healthy participants during the gesture task (first column in Fig. 2a). This similarity is less pronounced in the self-adaptor condition (second column in Fig. 2a). Furthermore, in both the no-hands (third column in Fig. 2a) and speech-only conditions (fourth column in Fig. 2a), the reproducibility maps are quite different among healthy participants. In contrast, stroke patients' reproducibility maps (Fig. 2b) vary dramatically from participant to participant irrespective of task. Variability of reproducibility maps seems less task dependent for stroke patients than for healthy participants. In the next section, we present a quantitative analysis to characterize this variability and compare the reliability difference between healthy and stroke groups and among the four different tasks.

### 3.2. Quantitative results

The fitted ROC curves for each of the four tasks are shown in Fig. 3a for healthy participants and in Fig. 3b for stroke patients. The estimated parameters, $A_z$, are shown in Table 2. An $F$-test was used to determine whether or not there were significant differences in variability between healthy participants and stroke patients, and among different tasks. After correction for multiple comparisons, the test results suggest the following:

- Inter-subject variability of healthy participants for the gesture task is significantly smaller than for the self-adaptor ($p < .05$), no-hands ($p < 10^{-3}$), or speech-only ($p < 10^{-3}$) tasks.
- There is no significant difference in variability among tasks for stroke patients.
- Inter-subject variability of stroke patients is significantly greater than healthy participants ($p < 5 \times 10^{-3}$) when they perform the gesture task, but there is no significant variability difference when they perform the other three tasks.

Maximum likelihood also provided an estimated proportion of truly active voxels, λ, in the brain for each individual subject. These results are represented in Fig. 4 for both healthy participants (Fig. 4a) and stroke patients (Fig. 4b). An *F*-test performed on the estimated proportion of active voxels demonstrates the following:

- There is no significant variability difference among tasks for healthy participants.
- Inter-subject variability of the stroke patients for the gesture task is significantly less than for the self-adaptor task (corrected $p < .05$), and marginally less than for the speech-only task (corrected $p = .08$).
- There is no significant difference in variability between healthy and stroke groups in the estimated proportion of active voxels.

To test whether there was any significant reliability difference for different tasks and different groups, a linear mixed effects model was used, with the fixed effects being those of the tasks (gesture, self-adaptor, no-gesture, and speech-only) and groups (healthy participants and stroke patients), and the random effects being the individual subjects. Since the variability test above suggests that there is a significant variability difference between the two groups and among some tasks, a weighted variability term is also included in the model. The test results suggest:

- Reliability for stroke patients is marginally worse than for healthy participants ($p = .046$).
- Reliability for the self-adaptor task is marginally less than for the gesture task (corrected $p = .055$). There are no other significant pair-wise reliability differences among tasks.
- Including a task by group interaction term does not significantly improve the model.

A similar linear mixed effects model was used on estimated λ and results suggest the following:

- There is no significant difference between healthy participants and stroke patients in the estimated proportion of activation.
- The estimated proportion of activation for the self-adaptor task is significantly greater than for the gesture task (corrected $p < .01$), the audio-only task (corrected $p < .005$), and marginally greater than the no-hands task (corrected $p = .08$).
- Again the task by group interaction term does not significantly improve the model.

## 4. Discussion

Our result suggests that the overall reliability in healthy participants is higher than in participants with brain injury. This is consistent with the result of Manoach et al. (2001) in patients with schizophrenia. In a working memory study, they found that the patients had less cross session reliability than healthy controls. Although there is evidence that BOLD activity disappears and reappears at different stages (time) of stroke (Binkofski & Seitz, 2004), such data cannot explain the current findings, since all of our patients are in the most chronic stage recovery (all more than 5 years poststroke). Clearly, in earlier stages, a lack of consistency of BOLD activity in stroke patients would clearly affect test reliability.

Reliability across time represents a measure within individuals. On this measure, the ecologically valid gesture task is more reliable than the quite unnatural self-adaptor task. This could result from the role of speech-associated (co-speech) gesture in language comprehension (McNeill, 1992). Such gestures, accompanying stories, could be used to improve language comprehension (Thompson & Massaro, 1986). By contrast, the self-adaptor are perceived as distracting. This could lead to less consistent brain activity across different experimental runs.

To the extent that attentional factors play a role in the difference between the two hand movement conditions, this result could be considered as extension of the results from Specht et al. (2003). They investigated modulation in the reliability of BOLD response by attentional effort, and suggested that greater attentional effort leads to more reliable results.

Whereas reliability pertains to individual subjects across time, variability measures the interchangeability of subjects within a group. In our study, the gesture task resulted in the least inter-subject variability compared with the other three tasks. There are several possible reasons for this. Again, the ecology of this task, with visual input from the face, mouth, and hands, should lead participants to better performance overall, including processes for attention, working memory, and comprehension (Goldin-Meadow, Kim, & Singer, 1999a; Goldin-Meadow & Sandhofer, 1999b; Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001; Sumby & Pollack, 1954; Summerfield, 1992). The gesture task thus may invoke similar (and well-exercised) natural mechanisms across individuals. It is thus possible that their performances are more similar on this task than on the others, whereas the other three tasks may be more dependent on participants. This is also consistent with the Specht et al. result (2003), where a task requiring less attentional effort was shown to have more variability that a task requiring a higher level of attention. The patients in the present study were heterogeneous in both stroke etiology and in their neural manifestations, including localization, and this would be expected to affect neural processing in diverse ways. It appears that this variability becomes the dominant factor in the stroke group, and leads to a relatively decreased contribution from other factors, such as task, in understanding overall variability. Within each stroke patient, the presence of a lesion leads to greater intra-individual variability in language performance (Howard, Patterson, Franklin, Morton, & Orchard-Lisle, 1984) and this likely dominates the reliability comparison, leading to less task modulation effects in reliability in the stroke group.

There are several limitations of our study. One of them is that since within-day and cross-day runs were pooled together to estimate the ROC curves, it is hard to know the degree to which each factor modulates reliability. Nonetheless, the four tasks were randomized and counterbalanced and we can assume that consistency within a session is greater than across sessions, and thus that reliability mainly depends on the cross-day factor. The second limitation is that because whole brain data were used to generate ROC curves, we do not know which specific regions in the brain play the critical roles in determining ROC curve performance. Since the ROC curve approach is not limited to the whole brain, the next step is to apply it to several language regions and compare reliability cross different regions in the brain. Another limitation comes from the methodology itself. When using this approach, we assume that all voxels in the brain are independent and have the same probability of activation. The first assumption is not valid because active voxels tend to be clustered together and probability of activation of a voxel is certainly task related. Maitra, Roys, and Gullapalli (2002) recommended an improved model by taking into account the spatial dependence among neighboring voxels and voxel-specific activation probability. However, the additional effort to estimate these parameters and resultant ROC curves did not seem to be different than those resulting from Genovese et al.'s method (1997).

The purpose of this study was to quantify how reliability is modulated by different tasks and different groups of participants. Before drawing any conclusions, we have to evaluate whether the reliability difference could be due to any other factors. In this study, all subjects were scanned by the same scanner and their data were processed using identical statistical methods, and thus these two possible confounds in assessment of reliability were minimized. In the experimental design, four different language tasks were randomized and counterbalanced, which helps to reduce the possibility that reliability differences among tasks might be due to the order of the stimuli. Since reliability could also be affected by the time interval between test and retest, a pair-wise *t*-test was performed between stroke and healthy groups, showing

no significant time difference. We also checked whether individual subject reliability depends on the time interval between test and retest and found that there was no such correlation for either group. It is commonly perceived that stroke patients have significant motion artifacts in fMRI images, and thus it is important to know if reliability differences across groups were affected by subject motion. To check this factor, we included motion regressors in the multiple linear regression and used this result to generate ROC curves, and found that this analysis did not improve the ROC performance for either healthy and stroke groups. This does not completely exclude the role of motion in affecting reliability, since such motion parameters are determined by automated algorithms that are not sensitive to all types of subject motion.

Although we used the total area under the curve as a single index to characterize the ROC curves, this measure might not provide a complete description of the curve, for example, two intersecting curves with the same area (Metz & Kronman, 1980). When high sensitivity or specificity is of more interest, partial area under the curve (Jiang, Metz, & Nishikawa, 1996) might be used to compare ROC curves. Another option is to use the mean difference between two populations, signal and noise, in the units of standard deviation of noise (“Parameter A”); and the ratio of standard deviations between noise and signal (“Parameter B”). Since these two parameters are highly correlated, statistical comparisons are more complicated and harder to interpret. Furthermore, the multivariate analysis can only test the hypothesis whether or not two curves are the same, but cannot determine which one is better, and thus is not a useful tool for our purpose. A single point can also be used to compare ROC curves, for example by comparing true positive rates at a fixed false positive rate (Genovese et al., 1997). In this case, the conclusion depends completely on the quality of estimation of a single point on the curve, and is consequently less accurate. In brief, area under the curve is a reasonably good index to serve our purpose.

## 5. Conclusions

By asking healthy participants and stroke patients to do exactly the same language tasks, we are able to compare the reliability, in terms of ROC curves, among tasks and between groups. Reliability here can be considered at two levels. Within-subject reliability is characterized by area under the curve, whereas across-subject reliability is characterized by the variability of areas under the curves. We find that stroke patients have less reliability compared with healthy participants, and that the gesture task is more reliable compared with the self-adaptor task. Inter-subject variability is more significant in the self-adaptor, speech-only, and self-adaptor tasks than in the gesture task. However, this variability difference was not observed for stroke patients.

## Acknowledgments

This research was supported by the National Institutes of Health under Grants P01-HD40605, R01-DC-3378, and R01-DC007488. The authors thank Charles Metz, Lorenzo Pesce, Howard Nusbaum, and Jeremy Skipper for helpful discussions.

## References

Aguirre GK, Zarahn E, D'Esposito M. The variability of human, BOLD hemodynamic responses. Neuroimage 1998;8:360–369. [PubMed: 9811554]

Binkofski F, Seitz RJ. Modulation of the BOLD-response in early recovery from sensorimotor stroke. Neurology 2004;63:1223–1229. [PubMed: 15477542]

Cao Y, D'Olhaberriague L, Vikingstad EM, Levine SR, Welch KM. Pilot study of functional MRI to assess cerebral activation of motor function after poststroke hemiparesis. Stroke 1998;29:112–122. [PubMed: 9445338]

Cohen MS, DuBois RM. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. Journal of Magnetic Resonance Imaging 1999;10:33–40. [PubMed: 10398975]

Constable RT, Skudlarski P, Gore JG. An ROC approach for evaluating functional brain MR imaging and postprocessing protocols. Magnetic Resonance in Medicine 1995;34:57–64. [PubMed: 7674899]

Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research 1996;29(3):162–173. [PubMed: 8812068]

Cox RW, Jesmanowicz A. Real-time 3D image registration for functional MRI. Magnetic Resonance in Medicine 1999;42(6):1014–1018. [PubMed: 10571921]

Cramer SC, Nelles G, Benson RR, Kaplan JD, Parker RA, Kwong KK, et al. A functional MRI study of subjects recovered from hemiparetic stroke. Stroke 1997;28:2518–2527. [PubMed: 9412643]

Fernández G, Specht K, Weis S, Tendolkar I, Reuber M, Fell J, et al. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 2003;60:969–975. [PubMed: 12654961]

Genovese CR, Noll DC, Eddy WF. Estimating test-retest reliability in functional MR imaging I: statistical methodology. Magnetic Resonance in Medicine 1997;38:497–507. [PubMed: 9339452]

Glover GH, Law CS. Spiral-in/out BOLD fMRI for increased SNR and reduced suspectibility artifacts. Magnetic Resonance in Medicine 2001;46:515–522. [PubMed: 11550244]

Goldin-Meadow S, Kim S, Singer M. What the teacher's hands tell the student's mind about math. Journal of Educational Psychology 1999a:720–730.

Goldin-Meadow S, Sandhofer CM. Gestures convey substantive information about a child's thoughts to ordinary listeners. Developmental Science 1999b:67–74.

Goldin-Meadow S, Nusbaum H, Kelly SD, Wagner S. Explaining math: gesturing lightens the load. Psychological Science 2001;12(6):516–522. [PubMed: 11760141]

Green, DM.; Swets, JA. Signal detection theory and psychophysics. Robert E. Krieger Publishing; Huntington, NY: 1976.

Howard, D.; Patterson, K.; Franklin, S.; Morton, J.; Orchard-Lisle, V. Variability and consistency in picture naming by aphasic patients. In: Rose, FC., editor. Recent advances in neurology. Raven Press; New York: 1984. p. 263-276.

Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996;201:745–750. [PubMed: 8939225]

LABROC4 is available on: www-radiology.uchicago.edu/krl/KRL_ROC/software_index.htm. Department of Radiology, Kurt Rossmann Laboratories for Radiologic Image Research, the University of Chicago, 5841 S. Maryland Avenue MC 2026, Chicago, IL 60637.

Le TH, Hu X. Methods for assessing accuracy and reliability in functional MRI. NMR in Biomedicine 1997;10:160–164. [PubMed: 9430342]

Liou M, Su H, Lee J, Cheng PE, Huang C, Tsai C. Bridging functional MR images and scientific inference: reproducibility maps. Journal of Cognitive Neuroscience 2003;15(7):935–945. [PubMed: 14628755]

Maitra R, Roys SR, Gullapalli RP. Test-retest reliability estimation of functional MRI data. Magnetic Resonance in Medicine 2002;48:62–70. [PubMed: 12111932]

Manoach DS, Halpern EF, Kramer TS, Chang Y, Goff DC, Rauch SL, et al. Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. American Journal of Psychiatry 2001;158:955–958. [PubMed: 11384907]

Marshall I, Simonotto E, Deary IJ, Maclullich A, Ebmeier KP, Rose EJ, et al. Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR imaging. Radiology 2002;233:868–877. [PubMed: 15498902]

McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frackowiak RSJ, Holmes AP. Variability in fMRI: an examination of intersession differences. NeuroImage 2000;11:708–734. [PubMed: 10860798]

McNeill, D. Hand and mind: What gestures reveal about thought. University of Chicago Press; Chicago, USA: 1992.

Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine 1978;8:283–298. [PubMed: 112681]

Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. Journal of Mathematical Psychology 1980;22:218–243.

Miezin FM, Maccotta L, Ollinger JM, Petersen SE, Buckner RL. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. NeuroImage 2000;11:735–759. [PubMed: 10860799]

Nandy RR, Cordes D. New approaches to receiver operating characteristic methods in functional magnetic resonance imaging with real data using repeated trials. Magnetic Resonance in Medicine 2004;52:1424–1431. [PubMed: 15562482]

Noll DC, Cohen JD, Meyer CH, Schneider W. Spiral K-space MRI of cortical activation. Journal of Magnetic Resonance Imaging 1995;5:49–56. [PubMed: 7696809]

Noll DC, Genovese CR, Nystrom LE, Vazquez AL, Forman SD, Eddy WF, et al. Estimating test-retest reliability in functional MR imaging II: application to motor and cognitive activation studies. Magnetic Resonance in Medicine 1997;38:508–517. [PubMed: 9339453]

Ogawa S, Tank DW, Menon R, Ellermann JM, Kim SG, Merkle H, et al. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proceedings of the National Academy of Sciences of the United States of America 1992;89:5951–5955. [PubMed: 1631079]

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 1971;9:97–113. [PubMed: 5146491]

Peck KK, Moore AB, Crosson BA, Gaiefsky M, Gopinath KS, White K, et al. Functional magnetic resonance imaging before and after aphasia therapy: shifts in hemodynamic time to peak during an overt language task. Stroke 2004;35:554–559. [PubMed: 14739418]

Pineiro R, Pendlebury S, Johansen-Berg H, Matthews PM. Altered hemodyamic responses in patients after subcortical stroke measured by functional MRI. Stroke 2002;33:103–109. [PubMed: 11779897]

Rutten GJM, Ramsey NF, Rijen PC, Veelen CWM. Reproducibility of fRMI-determined language lateralization in individual subjects. Brain and Language 2002;80:421–437. [PubMed: 11896650]

Sumby WH, Pollack I. Visual contribution of speech intelligibility in noise. The Journal of the Acoustical Society of America 1954;26(2):212–215.

Summerfield Q. Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London Series B—Biological Sciences 1992;335(1273):71–78.

Small SL, Flores D, Noll DC. Different neural circuits subserve reading before and after therapy for acquired dyslexia. Brain and Language 1998;62(2):298–308. [PubMed: 9576825]

Small SL, Hlustik P, Noll DC, Genovese C, Solodkin A. Cerebellar hemispheric activation ipsilateral to the paretic hand correlates with functional recovery after stroke. Brain 2002;125(Pt 7):1544–1557. [PubMed: 12077004]

Specht K, Willmes K, Shah NJ, Jäncke L. Assessment of Reliability in functional imaging studies. Journal of Magnetic Resonance Imaging 2003;17:463–471. [PubMed: 12655586]

Thompson LA, Massaro DW. Evaluation and integration of speech and pointing gestures during referential understanding. Journal of Experimental Child Psychology 1986;42:144–168. [PubMed: 3772293]

Waldvogel D, Gelderen PV, Immisch L, Pfeiffer C, Hallett M. The variability of serial fMRI data: correlation between a visual and a motor task. NeuroReport 2000;11:3843–3847. [PubMed: 11117501]

Wei X, Yoo SS, Dickey CC, Zou KH, Guttmann CR, Panych LP. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. NeuroImage 2004;21:1000–1008. [PubMed: 15006667]

Xiong J, Rao S, Jerabek P, Zamarripa F, Woldprff M, Lancaster J, et al. Intersubject variability in cortical activations during a complex language task. NeuroImage 2000;12:326–339. [PubMed: 10944415]

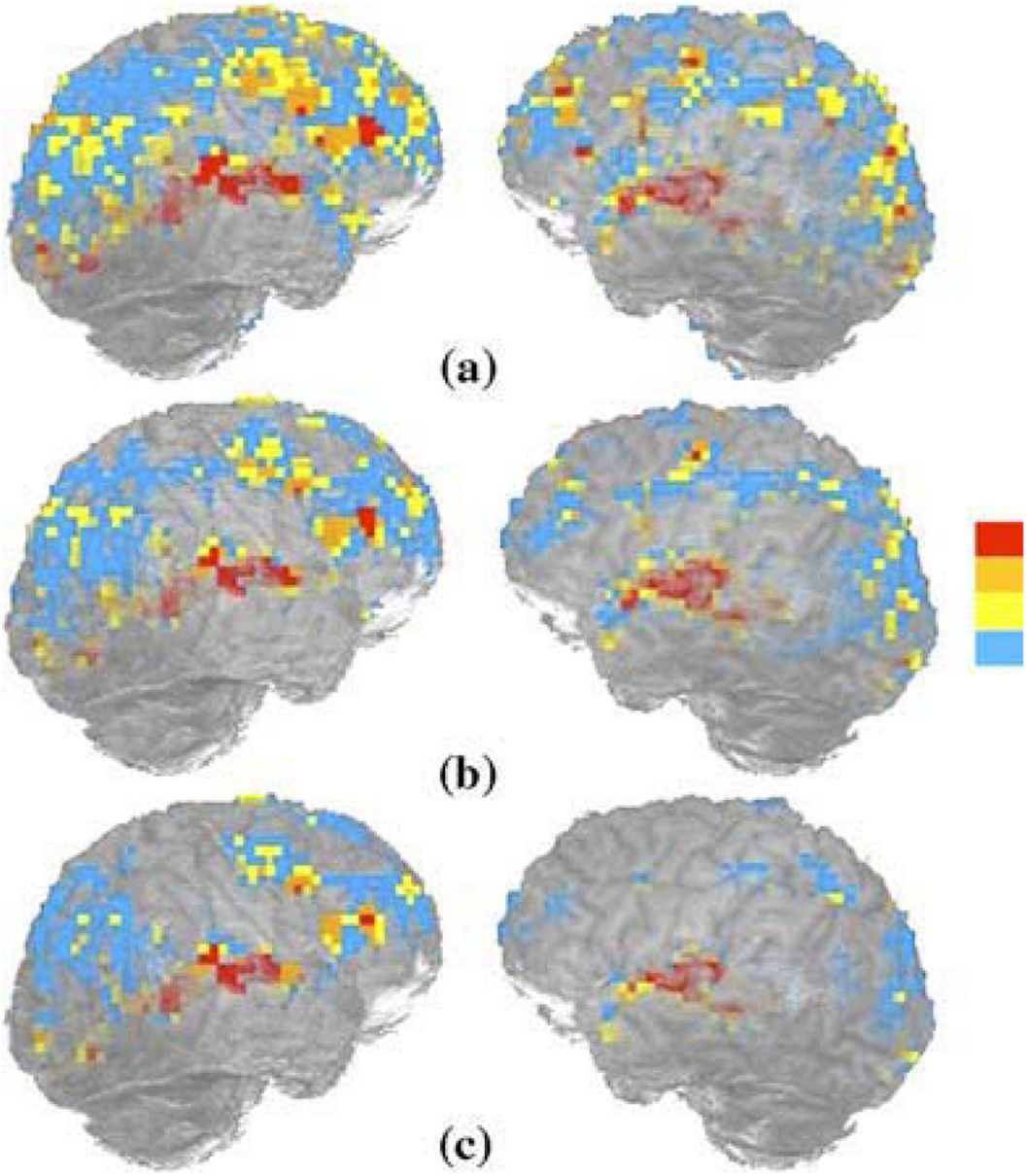


**Fig. 1.**

An illustration of reproducibility maps of participant 005 performing gesture task at three different thresholds, uncorrected $p < 10^{-2}$ (a), $p < 10^{-3}$ (b). and $p < 10^{-4}$ (c). The color bar changing from blue to red corresponds to a voxel classified as active one to four times.

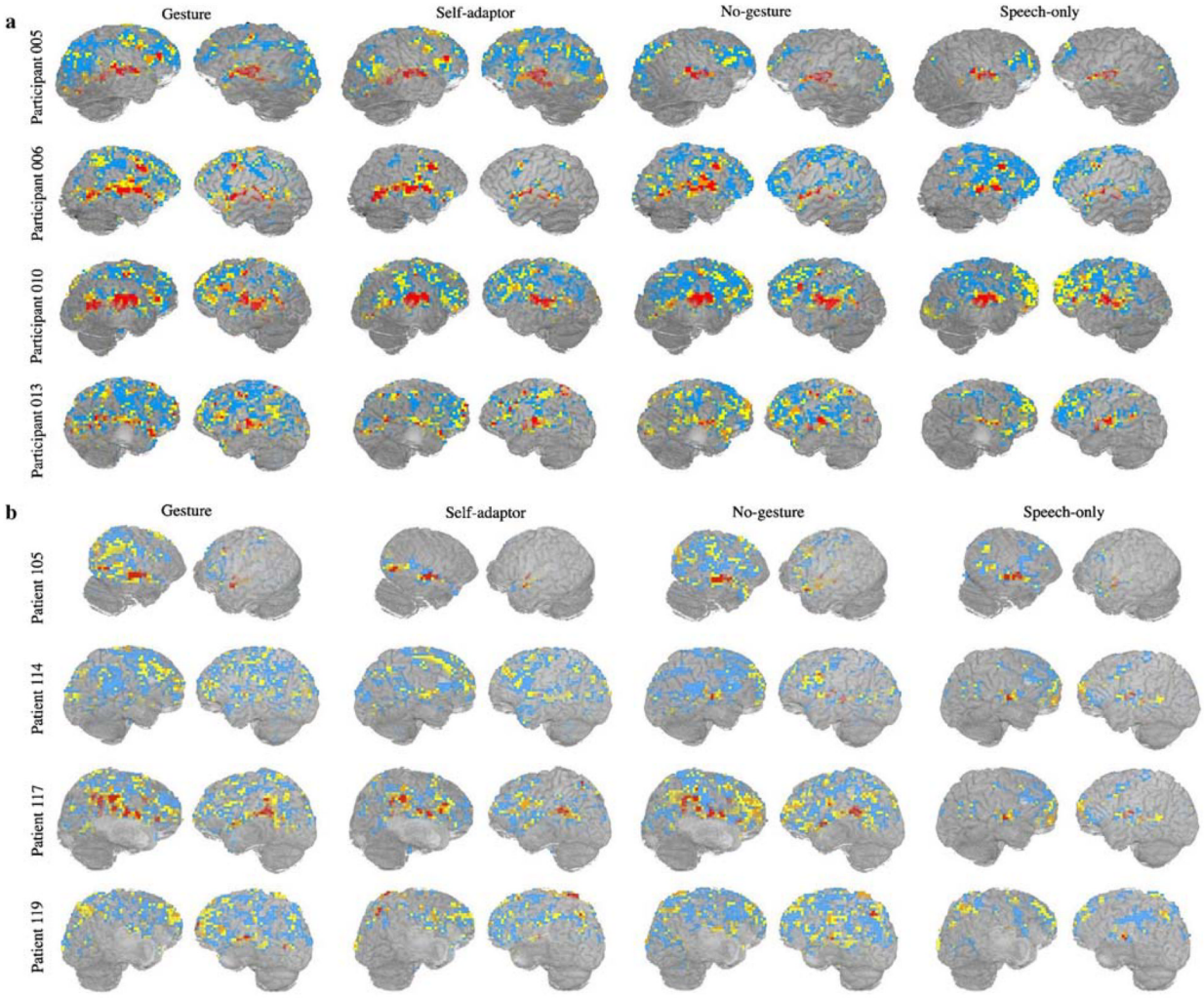


**Fig. 2.**

(a) Representative reliability maps for four healthy participants at a fixed threshold (uncorrected $p < 10^{-3}$). Color coding is the same as in Fig. 1. (b) Representative reliability maps for four stroke patients at a fixed threshold (uncorrected $p < 10^{-3}$). Color coding is the same as in Fig. 1.

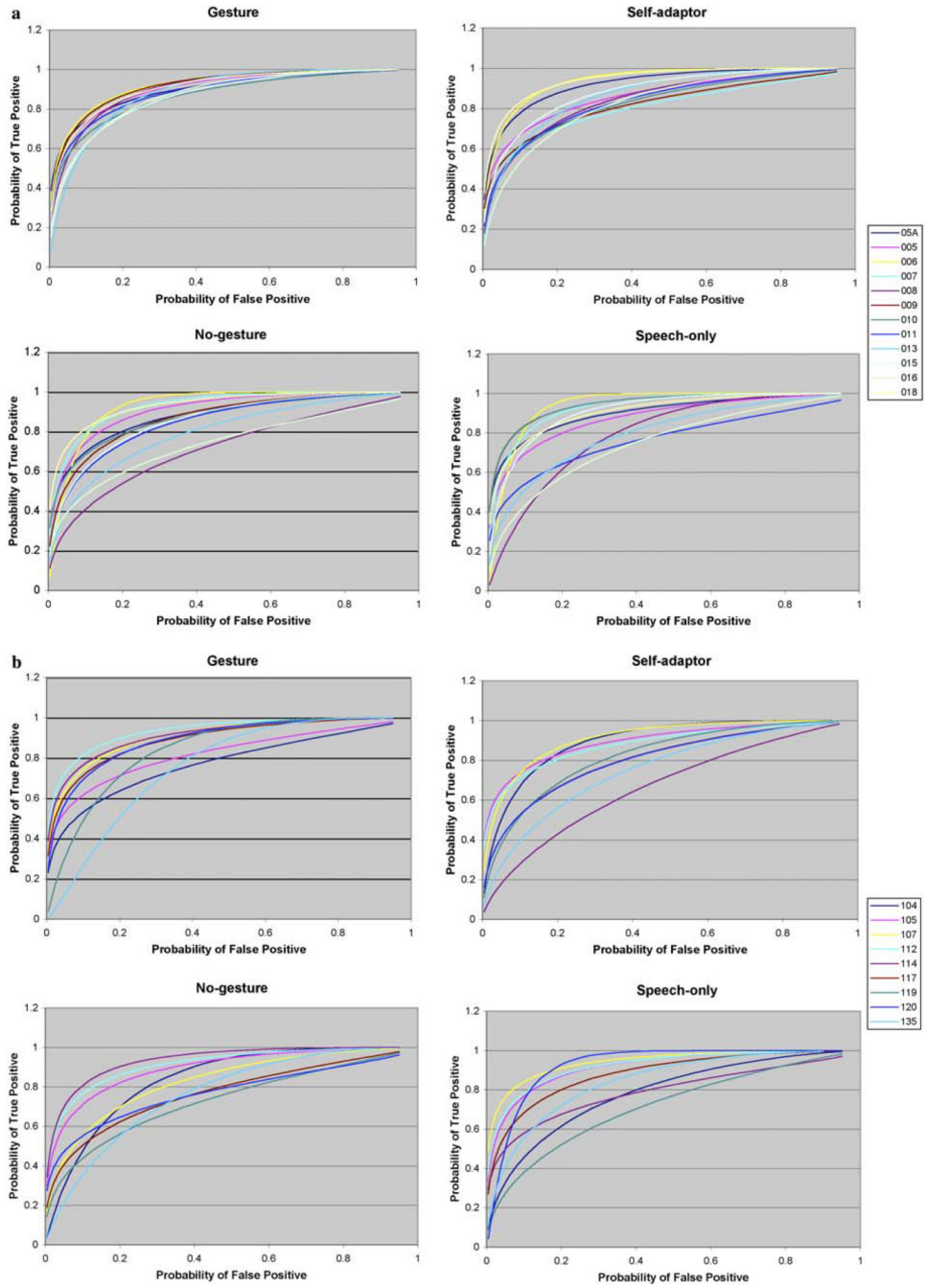


**Fig. 3.**

(a) ROC curves of 12 healthy participants conducting gesture, self-adaptor, no-gesture, and speech-only tasks. (b) ROC curves of nine stroke patients conducting gesture, self-adaptor, no-gesture, and speech-only tasks.

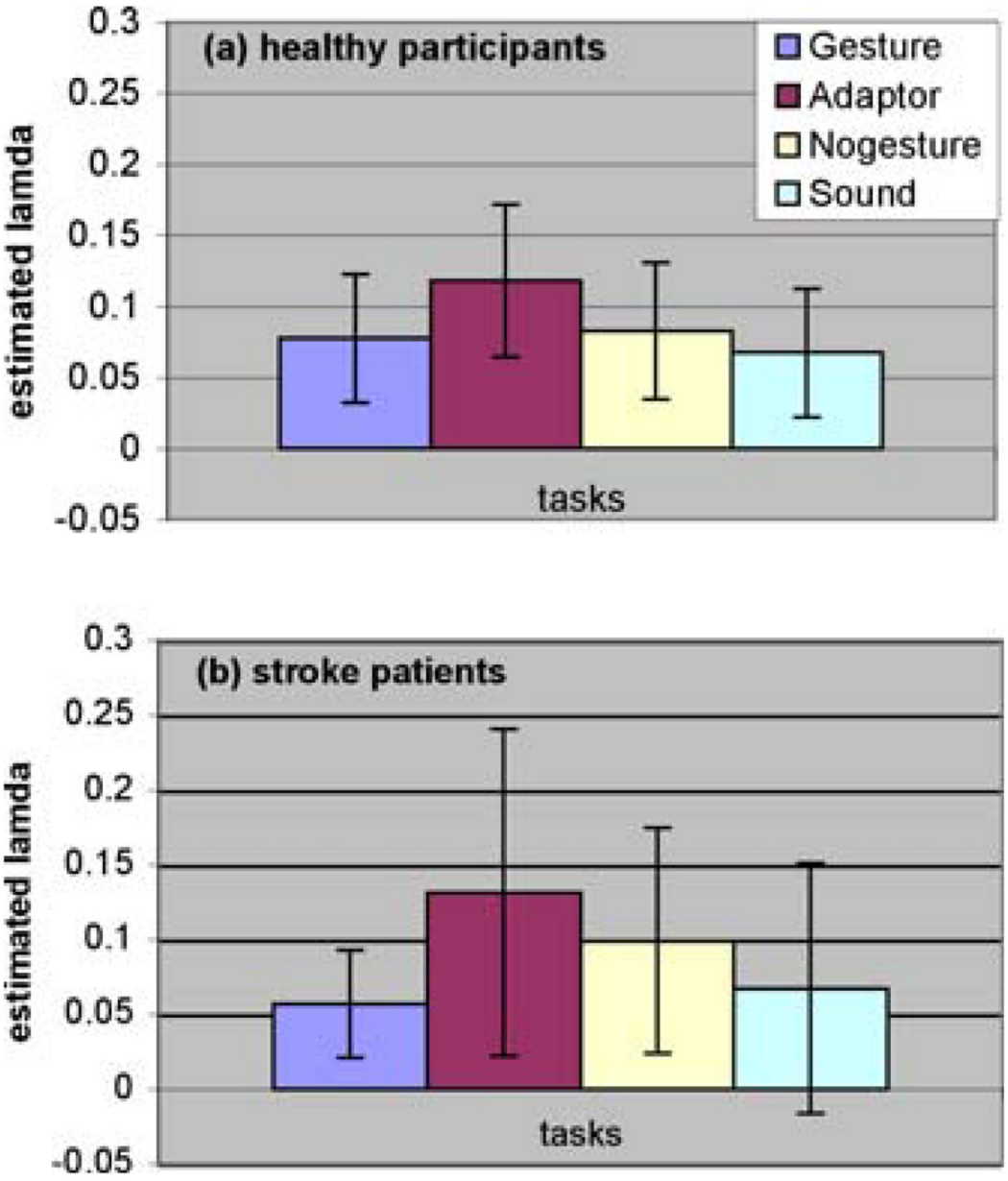


**Fig. 4.**
Estimated proportion of active voxels for healthy participants (a) and stroke patients (b).

**Table 1**

Time interval between test and retest for healthy subjects and stroke patients

| Time interval | Healthy participants | Days | Stroke patients | Days |
|---|---|---|---|---|
| Within a week | sub010 | 5 | sub112 | 1 |
| | sub008 | 6 | sub114 | 1 |
| | | | sub135 | 1 |
| | | | sub105 | 3 |
| Within a month | sub018 | 11 | | |
| | sub009 | 14 | | |
| | sub015 | 14 | | |
| | sub013 | 14 | sub120 | 17 |
| | sub011 | 28 | sub119 | 27 |
| More than a month | sub006 | 42 | | |
| | sub007 | 45 | | |
| | sub016 | 56 | sub104 | 61 |
| | sub005 | 88 | sub117 | 63 |
| | sub05A | 199 | sub107 | 343 |

**Table 2**

Area under the curve ($A_z$) for healthy participants (a) and stroke patients (b) and corresponding means and standard deviations for gesture, self-adaptor, no-gesture and speech-only conditions

| | Gesture | Self-adaptor | No-gesture | Speech-only |
|---|---|---|---|---|
| *(a) Healthy participants* | | | | |
| Sub05A | 0.90 | 0.92 | 0.88 | 0.90 |
| Sub005 | 0.91 | 0.86 | 0.91 | 0.88 |
| Sub006 | 0.92 | 0.94 | 0.93 | 0.94 |
| Sub007 | 0.91 | 0.81 | 0.93 | 0.93 |
| Sub008 | 0.90 | 0.85 | 0.72 | 0.80 |
| Sub009 | 0.92 | 0.82 | 0.87 | 0.94 |
| Sub010 | 0.87 | 0.83 | 0.87 | 0.94 |
| Sub011 | 0.89 | 0.84 | 0.85 | 0.77 |
| Sub013 | 0.88 | 0.87 | 0.80 | 0.80 |
| Sub015 | 0.87 | 0.88 | 0.86 | 0.91 |
| Sub016 | 0.87 | 0.83 | 0.74 | 0.76 |
| Sub018 | 0.91 | 0.94 | 0.94 | 0.91 |
| Mean | 0.90 | 0.87 | 0.86 | 0.87 |
| *SD* | 0.02 | 0.05 | 0.07 | 0.07 |
| *(b) Stroke patients* | | | | |
| Sub104 | 0.77 | 0.90 | 0.84 | 0.78 |
| Sub105 | 0.82 | 0.90 | 0.89 | 0.92 |
| Sub017 | 0.91 | 0.91 | 0.83 | 0.94 |
| Sub112 | 0.93 | 0.88 | 0.91 | 0.92 |
| Sub114 | 0.91 | 0.67 | 0.94 | 0.79 |
| Sub117 | 0.89 | 0.80 | 0.77 | 0.88 |
| Sub119 | 0.84 | 0.82 | 0.73 | 0.72 |
| Sub120 | 0.89 | 0.80 | 0.77 | 0.93 |
| Sub135 | 0.76 | 0.75 | 0.76 | 0.84 |
| Mean | 0.86 | 0.83 | 0.83 | 0.86 |
| *SD* | 0.06 | 0.08 | 0.07 | 0.08 |