Approaching the Ground Truth: Revealing the Functional Organization of Human Multisensory STC Using Ultra-High Field fMRI

Francesco Gentile,¹ Nienke van Atteveldt,^{3,4,5} Federico De Martino,² and Rainer Goebel^{1,2}

¹Netherlands Institute for Neuroscience, 1105 BA Amsterdam, The Netherlands, ²Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6229 ER Maastricht, The Netherlands, ³Department of Clinical Developmental Psychology, Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands, ⁴LEARN! Research Institute for Learning and Education, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands, and ⁵Institute of Brain and Behavior Amsterdam, Faculty of Behavioral and Movement Sciences, VU, 1081 HV Amsterdam, The Netherlands

Integrating inputs across sensory systems is a property of the brain that is vitally important in everyday life. More than two decades of fMRI research have revealed crucial insights on multisensory processing, yet the multisensory operations at the neuronal level in humans have remained largely unknown. Understanding the fine-scale spatial organization of multisensory brain regions is fundamental to shed light on their neuronal operations. Monkey electrophysiology revealed that the bimodal superior temporal cortex (bSTC) is topographically organized according to the modality preference (visual, auditory, and bimodal) of its neurons. In line with invasive studies, a previous 3 Tesla fMRI study suggests that the human bSTC is also topographically organized according to modality preference (visual, auditory, and bimodal) when analyzed at $1.6 \times 1.6 \times 1.6 \text{ mm}^3$ voxel resolution. However, it is still unclear whether this resolution is able to unveil an accurate spatial organization of the human bSTC. This issue was addressed in the present study by investigating the spatial organization of functional responses of the bSTC in 10 participants (from both sexes) at $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ and $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ using ultra-high field fMRI (at 7 Tesla). Relative to $1.5 \times 1.5 \times 1.5 \text{ mm}^3$, the bSTC at $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ resolution was characterized by a larger selectivity for visual and auditory modalities, stronger integrative responses in bimodal voxels, and it was organized in more distinct functional clusters indicating a more precise separation of underlying neuronal clusters. Our findings indicate that increasing the spatial resolution may be necessary and sufficient to achieve a more accurate functional topography of human multisensory integration.

Key words: bimodal integration index; bimodal STC; category selectivity; multisensory integration; ultra-high field fMRI

Significance Statement

The bimodal superior temporal cortex (bSTC) is a brain region that plays a crucial role in the integration of visual and auditory inputs. The aim of the present study was to investigate the fine-scale spatial organization of the bSTC by using ultra-high magnetic field fMRI at 7 Tesla. Mapping the functional topography of bSTC at a resolution of $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ revealed more accurate representations than at lower resolutions. This result indicates that standard-resolution fMRI may lead to wrong conclusions about the functional organization of the bSTC, whereas high spatial resolution is essential to more accurately approach neuronal operations of human multisensory integration.

Introduction

Multisensory integration is an essential integrative operation of the brain. It consists of combining the pieces of information about an object or event that are conveyed through distinct sen-

sory systems. In nonhuman primates, a brain region that has repeatedly been found to be involved in multisensory integration is the upper bank of the superior temporal sulcus (STS) (Seltzer et al., 1996; Dahl et al., 2009). Crucially, it has been observed that neurons responding to a specific modality (visual or auditory)

Received Jan. 11, 2017; revised June 22, 2017; accepted Aug. 5, 2017.

Author contributions: F.G., N.v.A., and R.G. designed research; F.G., N.v.A., and F.d.M. performed research; R.G. contributed unpublished reagents/analytic tools; F.G. analyzed data; F.G., N.v.A., F.d.M., and R.G. wrote the paper.

This work was supported by the 2014-MBIC funding from the Faculty of Psychology and Neuroscience of Maastricht University, European Research Council Grant 269853, and European FET Flagship Human Brain Project FP7-ICT-2013-FET-F/604102.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Francesco Gentile, Netherlands Institute for Neuroscience, Meibergdreef 47, 1105 BA Amsterdam, The Netherlands. E-mail: f.gentile@nin.knaw.nl.

DOI:10.1523/JNEUROSCI.0146-17.2017

Copyright © 2017 the authors 0270-6474/17/3710104-10\$15.00/0

tend to cluster together, or cluster with bimodal neurons, and are spatially separated from the other modality-specific neurons (Schroeder et al., 2003; Wallace et al., 2004; Schroeder and Foxe, 2005; Dahl et al., 2009).

Using conventional voxel resolution, typically $\sim 3 \times 3 \times 3 \text{ mm}^3$, human fMRI studies have identified a portion of the superior temporal cortex (STC), the so-called bimodal STC (bSTC), which responds to both visual and auditory inputs, and where responses are enhanced for audiovisual inputs compared with when each input is presented alone (Calvert et al., 2001; Beauchamp et al., 2004a; van Atteveldt et al., 2004; Amedi et al., 2005; Beauchamp, 2005b; Doehrmann and Naumer, 2008; Werner and Noppeney, 2010; van Atteveldt et al., 2010). However, at $3 \times 3 \times 3 \text{ mm}^3$ resolution, the voxel size is too large to interpret the observed activation in the bSTC at a more fine-grained neuroanatomically meaningful level. Indeed, the enhanced fMRI response to bimodal stimulation observed within a conventional bSTC voxel can reflect bimodal neuronal populations, but it can also be the result of pooled neuronal responses from separate unimodal (visual, auditory) populations or even the result of a mixture of unimodal and bimodal neuronal populations (Beauchamp, 2005a; Laurienti et al., 2005; Goebel and van Atteveldt, 2009).

In a first attempt to disentangle these alternatives, Beauchamp et al. (2004b) increased the resolution of the fMRI measurement to $1.6 \times 1.6 \times 1.6 \text{ mm}^3$ (middle-res) using parallel imaging at 3 Tesla (3T), and provided the first evidence that the human bSTC is composed of unimodal (visual and auditory) and bimodal regions orderly organized in a “patchy” fashion. However, it is still unclear to what extent the topographical organization of the human bSTC revealed in this study is accurate in terms of neural representation of visual, auditory, and bimodal subregions within the STC. This issue was addressed in a monkey electrophysiological study where single-unit data were spatially smoothed to simulate the expected results of lower resolution (e.g., fMRI) investigations (Dahl et al., 2009). It was observed that, at $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ resolution, the amount of bimodal voxels was underestimated compared with the unsmoothed (single and multiunit) level resolution, but that, when increasing the resolution to $0.75 \times 0.75 \times 0.75 \text{ mm}^3$, the proportion of bimodal, visual, and auditory voxels was similar to the topography observed at the (single and multiunit) level. These results suggest that the previously described topographical organization of the human bSTC at middle-res (Beauchamp et al., 2004b) might not have captured the underlying neuronal organization and that to approach the true spatial and functional organization of human bSTC, the resolution needs to be increased.

In the present fMRI study, we increased resolution up to $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ by using ultra-high magnetic field fMRI at 7 Tesla (7T). We expected that the smaller voxel size would lead to responses that are more specific to the underlying modality-related neuronal populations, resulting in a more accurate mapping of the human bSTC. This hypothesis was tested by comparing various properties of unimodal (visual and auditory) and bimodal (audiovisual) clusters in the bSTC at $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ (middle-res) and $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ (high-res). The bSTC was independently localized via a $3.5 \times 3.5 \times 3.5 \text{ mm}^3$ dataset (low-res) acquired at 3T. Statistics on low-, middle-, and high-resolution datasets were performed at the single-subject level and in native space. By doing so, we avoided any blurring induced by multisubject analysis in a normalized anatomical space and, instead, we preserved the functional specificity of each dataset.

Materials and Methods

Subjects. Eleven healthy volunteers (7 female, age 25.5 ± 3.5 years) participated in the fMRI experiment that consisted of three experimental sessions: one at 3T and two at 7T. We excluded the data from one participant from the statistical analysis as the head-movement inside the scanner was significantly suprathreshold (i.e., larger than a single voxel). Participants were all right-handed and native Dutch speakers, and they had normal or corrected-to-normal vision. The study and its procedures were approved by the Ethical Review Committee Psychology and Neuroscience of Maastricht University.

Stimuli. Unimodal (visual and auditory) and bimodal activation within the STC were elicited by using a well-tested experimental design consisting of single letters presented one at the time in a visual and auditory mode (unimodal) or presented simultaneously in both modalities (bimodal congruent/incongruent; Fig. 1) (van Atteveldt et al., 2004). The unimodal auditory stimuli consisted of speech sounds that were digitally recorded (sampling rate 44.1 kHz, 16 bit quantization) from a female speaker. Recordings were bandpass filtered (180–10,000 Hz) and resampled at 22.05 kHz. The speech sounds selected for the fMRI experiment were only those that were recognized with 100% accuracy in a pilot experiment ($n = 10$). The final set of selected consonants consisted of b, d, g, h, k, l, n, p, r, s, t, and z, and the selected vowels were a, e, i, y, o, and u. The average duration of the speech sounds was 352 ± 5 ms, and the average sound intensity level was 71.3 ± 0.2 dB. The unimodal visual stimuli consisted of lowercase letters in Arial font corresponding to the 18 selected speech sounds. A single letter covered a visual angle of $2.23^\circ \times 2.23^\circ$ and was presented in white in the center of a black background for 332 ms. During fixation periods, a white fixation cross was presented in the center of the screen.

Experimental design. The experiment consisted of three sessions: the first acquired at 3T and the second and third at 7T. A total of four experimental runs were presented in each session using a silent sparse sampling fMRI paradigm. Within a run, the four different conditions, namely, unimodal visual, unimodal auditory, bimodal congruent, and bimodal incongruent, were presented in blocks. A run started with 30 s of rest followed by a total of 12 stimulation blocks (3 for each condition) alternating with 12 resting periods each lasting 16 s. The rest period following the last stimulation block lasted for 24 s. Each stimulation block lasted 24 s and consisted of 6 mini-blocks (Fig. 1a). The total duration of a single mini-block was 4 s during which stimuli were presented in silence for 2 s and one functional volume was acquired in a resting period for the remaining 2 s (i.e., silent-delay = 2000 ms, acquisition time $TA = 2000$ ms, volume $TR = 4000$ ms). In a mini-block, a sequence of three letters were presented visually, aurally, or in both modalities. To avoid that a sequence of three letters presented in a mini-block would form a word containing any meaningful content, the three letters consisted of only vowels or consonants that were chosen randomly from the set of available stimuli. In the bimodal conditions, visual letters and sounds were presented simultaneously (Fig. 1b).

Acquisition procedure and preprocessing of the low-res dataset at 3T. The acquisition of the low-res dataset was performed using a whole-body 3T Magnetom Prisma scanner (Siemens Medical System) equipped with a 64-channel head-neck coil. Thirty-three oblique axial slices (in-plane resolution: $3.5 \text{ mm} \times 3.5 \text{ mm}$, slice thickness: 3.5 mm, no gap) covering the entire brain were acquired using an echo planar imaging sequence ($TA = 2000$ ms, silent-delay = 2000 ms, $TR = 4000$ ms, $TE = 28$ ms, nominal $FA = 45^\circ$, matrix: 64×64 , FOV: $224 \text{ mm} \times 224 \text{ mm}$). We acquired a total of 4 runs consisting of 128 volumes each. Functional slices were aligned to a high-resolution 3D anatomical dataset (T1-weighted data) acquired in the middle of the entire session and consisting of 192 slices (ADNI sequence: $TR = 2250$ ms, $TE = 2.21$ ms, $FA = 9^\circ$, voxel dimension = $1 \times 1 \times 1 \text{ mm}^3$). Functional and anatomical data were analyzed using the BrainVoyager QX software package (version 2.8.4.2645, 64-bit, Brain Innovation). The preprocessing procedure of the functional datasets consisted of three steps: correcting for slice scan-time differences, motion artifacts correction in three dimensions, and high-pass filtering using a GLM Fourier basis set to remove slow-frequency drifts up to 2 cycles per time course, but no spatial smoothing.

Figure 1. *a*, Experimental design. *b*, Stimuli.

After preprocessing, functional data were coregistered to the individual anatomical T1 scan using the position information recorded at time of acquisition (initial alignment) followed by fine alignment. In the fine alignment procedure, the source was aligned to the target (T1 anatomical scan) using the “normalized gradient field” algorithm (implemented in BrainVoyager). More specifically, when intensity changes in the same way in both volumes at corresponding positions, the normalized gradient field considers source and target volume as aligned. Differently from procedures that use the intensity per se, the normalized gradient field procedure is largely unaffected by inhomogeneities in the data. For coregistration, the functional run following the acquisition of the T1 anatomical scan was used. Alignment across functional runs was performed within the motion artifacts correction algorithm by using the same target volume for all volumes of all runs. The final version of the functional data consisted of a 4-dimensional (x, y, z, t) dataset for each run and participant resampled using trilinear interpolation in the space of the anatomical scan.

Acquisition procedure and preprocessing of the middle-res/high-res dataset at 7T. The acquisition at 7T consisted of two sessions where 4 runs at middle-res (first session) and 4 runs at high-res (second session) were acquired in a whole-body 7T Magnetom scanner (Siemens Medical System) equipped with a 32-channel head coil. For each of the 8 runs (4 at middle-res and 4 at high-res), a total of 128 volumes were acquired using a standard single shot gradient echo EPI sequence. Slices were placed to cover the previously functionally localized region of interest (ROI). In the middle-res session, 35 oblique axial slices (in-plane resolution: 1.5 mm × 1.5 mm, slice thickness: 1.5 mm, no gap, TA = 2000 ms, silent-delay = 2000 ms, TR = 4000 ms, TE = 19 ms, nominal FA = 80°, matrix: 128 × 128, FOV: 192 mm × 192 mm) were acquired. In the high-res session, we acquired 33 slices (in-plane resolution: 1.1 mm × 1.1 mm, slice thickness: 1.1 mm, no gap, TA = 2000 ms, silent-delay = 2000 ms, TR = 4000 ms, TE = 19 ms, nominal FA = 86°, matrix: 174 × 174, FOV: 192 mm × 192 mm). The phase encoding direction in both sessions was from anterior to posterior (A >> P). To correct for geometric distortions of the EPI images (see below), we also acquired five volumes in the opposite direction, namely, from posterior to anterior (P >> A) before the acquisition of each dataset. At the beginning of the first ses-

sion, T1-weighted data (MPRAGE sequence: TR = 3100 ms; TE = 2.52 ms; FA = 5°, voxel dimension = 0.6 × 0.6 × 0.6 mm³) and proton density-weighted data (MPRAGE sequence: TR = 1440 ms; TE = 2.52 ms; flip angle = 5°, voxel dimension = 0.6 × 0.6 × 0.6 mm³) were acquired. Functional and anatomical middle-res and high-res datasets were preprocessed in the same way as the low-res dataset (using the BrainVoyager QX software package) except for two additional steps: (1) anatomical images were corrected for bias field inhomogeneities by computing the ratio with the simultaneously obtained proton density-weighted images (Van de Moortele et al., 2009) followed by an intensity inhomogeneity correction step to remove residual inhomogeneities; and (2) EPI distortion correction was performed on the functional volumes using the COPE (Correction based on Opposite Phase Encoding) plugin in BrainVoyager (Heinrich et al., 2014). After distortion correction, coregistration of the 7T data followed a procedure similar to the one used for the dataset acquired at 3T. The only difference consisted of aligning all single runs individually to the anatomical reference and using within rather than across-run motion correction. An ROI GLM analysis was performed using the middle-res and high-res datasets considering as ROI the bSTC localized using the low-res dataset. To do so, the previously defined bSTC was coregistered to the functional middle-res and high-res datasets using as information the alignment across the anatomical datasets of the 3T and 7T sessions.

Statistical analyses. The low-res dataset was used to localize the bSTC for each subject independently. A standard GLM analysis was performed with the four stimulus types (i.e., unimodal visual, unimodal auditory, bimodal congruent, and bimodal incongruent) as predictors. The bSTC was defined as the regions that responded to both auditory (vs baseline) and visual (vs baseline) unimodal conditions (corrected for multiple comparison via the false discovery rate approach: FDR = $q(\text{FDR}) < 0.05$ (Fig. 2) (Genovese et al., 2002).

The ROI-GLM used to analyze the middle-res and high-res datasets was similar to the GLM performed at 3T. The same four stimulus types were defined as predictors in the GLM to identify visual, auditory, and bimodal clusters within each subject’s localized bSTC, per hemisphere. More specifically, the bimodal clusters consisted of voxels significantly responding to both visual and auditory conditions. The visual voxels

Figure 2. Definition of the bSTC at $3.5 \times 3.5 \times 3.5 \text{ mm}^3$ (low-res) acquired at 3T. Analysis and results. Statistical conjunction used to define the bSTC. The bSTC localized for all 10 subjects.

Figure 3. Definition of the auditory, visual, and bimodal regions within the bSTC at $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ (middle-res) and at $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ (high-res) acquired at 7T. Analysis and results. **a**, Statistical definition of visual, auditory, and bimodal regions localized at middle-res and high-res within the bSTC. Clusters of voxels that exhibited substantial activation levels ($*p < 0.05$, uncorrected) were selected for further analysis. **b**, Example of activation maps (Subject 02) at middle-res and high-res. **c**, Macro-analysis. Proportion of voxels within the visual, auditory, and bimodal regions at low-res, middle-res, and high-res.

were those activated in response to the visual condition (vs baseline) and responding to the visual condition in a significantly larger extent than the auditory condition. Similarly, auditory clusters consisted of voxels significantly responding to the auditory condition (vs baseline) and whose response was significantly larger for the auditory than visual condition

(Fig. 3a). The significance threshold to define auditory, visual, and bimodal clusters was set to a p value < 0.05 (uncorrected) as in the study of Beauchamp et al. (2004b), to enable comparing our results to their findings. To ensure reliability, the localization of visual, auditory, and bimodal voxel was further validated by performing a test-retest spatial

Figure 4. Selectivity index analysis. **a**, Example of activation maps (Subject 02) at middle-res and high-res. **b**, Computational procedure of the unimodal selectivity index. Schematic representation (in terms of time course averaged over repetitions) of the β weights from the GLM analysis related to the visual and auditory condition and extracted from the visual (left) and auditory (right) clusters within the bSTC at middle-res and high-res. The visual selectivity index at middle-res and high-res was computed as the difference (indicated with an arrow) between the visual-related β and the auditory-related β extracted from the visual clusters at middle-res and high-res, respectively. The auditory selectivity index was computed in a similar way by just inverting the two variables of the subtraction. **c**, Visual (left) and auditory (right) selectivity indexes for each subject for middle-res and high-res.

reliability analysis (Beauchamp et al., 2004b) separately for middle-res and high-res datasets. More specifically, the same ROI-GLM that was used to define unimodal and bimodal clusters was performed twice for each resolution: one time using the odd and one time using the even runs. As a measure of the consistency between the two splits, we computed the spatial cross-correlation between the maps of visual, auditory, and bimodal voxels across the two splits, separately for the middle-res and high-res datasets. First, we calculated for each split the t statistic of the contrasts used to define visual, auditory, and bimodal clusters for all voxels within bSTC (localized at low-res). Therefore, a t statistic related to visual, auditory, and bimodal was associated to each voxel. A voxelwise cross-correlation between those t statistics of the two splits was then computed for each resolution. The test-retest reliability analysis was the only analysis in which we split the 4 runs in odd and even. All other analyses described in the paper and figures are performed on all 4 runs.

We then more specifically analyzed the functional properties of the visual, auditory, and bimodal clusters. First, the three voxel types were analyzed at a macroscopic level by computing the proportion of voxels of each type present in the entire bSTC-ROI (Dahl et al., 2009) and comparing these proportions statistically within and across resolutions. Because it was not possible to assess the distribution of our data due to the small sample size ($n = 10$) we used, for this analysis, nonparametric tests (Kruskal–Wallis test with resolution and voxels' type as factors).

Next, we tested whether increasing the resolution from middle-res to high-res resulted in a more accurate unimodal (visual and auditory) and bimodal spatial map of the bSTC. To do so, the three voxel types at middle-res and high-res were characterized in terms of (1) voxel-type selectivity for unimodal (visual and auditory) and bimodal conditions and (2) spatial organization. The selectivity analysis for unimodal clusters consisted of computing and comparing the visual and auditory selectivity index, in the visual and auditory clusters, respectively, within the bSTC at middle-res and high-res. The visual selectivity index of the visual clusters was computed by subtracting the β estimates (calculated from

the bSTC-GLM and averaged over the visual voxels) of the visual predictor from the β estimates of the auditory predictor. The auditory selectivity was computed by the opposite subtraction ($\beta_{\text{auditory}} - \beta_{\text{visual}}$) within the auditory clusters (Fig. 4a,b). Because of the way bimodal voxels were defined (Fig. 3a), it was not possible to compute selectivity for bimodal clusters in the same way as visual and auditory clusters. Instead, it was computed based on two different measurements. The first was the proportion of equally responsive bimodal voxels defined as those bimodal voxels whose visual and auditory responses were statistically indistinguishable. To fully understand the valence of this measurement and the way we used to compute it, it is important to remind that the bimodal voxels, described so far, were defined as those voxels responding to both the visual and auditory conditions. Therefore, the bimodal voxels could, in principle, overlap with visual or auditory voxels (i.e., bimodal voxels could be at the same time visual or auditory voxels). The extent of such an overlap was inversely proportional to the amount of equally responsive bimodal voxels (i.e., the smaller the overlap the larger the proportion; Fig. 5a). Therefore, to compute the proportion of equally responsive bimodal voxels, we calculated that overlap. We proceeded by determining the number of visual and auditory voxels overlapping with the bimodal voxels and then we divided the obtained values by the number of visual voxels (composing the visual clusters) and auditory voxels (composing the auditory clusters), respectively. The second parameter we chose to evaluate the selectivity of bimodal voxels was their integrative power. We assessed to what extent unimodal visual and auditory stimuli were integrated within the bimodal clusters (integrative power of the bimodal clusters) by calculating the bimodal integrative index according to the max criterion: $[\text{bimodCo} - \max(\text{aud}, \text{vis})] / \max(\text{aud}, \text{vis})$ (Fig. 6a) (Beauchamp, 2005a; Goebel and van Atteveldt, 2009).

Finally, we investigated the spatial organization of the visual, auditory, and bimodal clusters (within the bSTC) at high-res and middle-res. More specifically, the position of the centroid of all clusters types was computed; for each of them, the surrounding clusters (at various distances in

Figure 5. Overlap analysis. **a**, Computational procedure. Left, Schematic representation (i.e., both the boxes and arrows are not quantitatively scaled) illustrating the spatial overlap at middle-res between the visual and auditory clusters (green box and red box, respectively) with the bimodal cluster (blue box). Within each cluster/box, we also draw some small boxes representing voxels in that cluster. The visual-with-bimodal overlap was computed by identifying the amount of voxels that were present in both the visual and bimodal clusters (green dashed area) divided by the amount of visual voxels composing the visual clusters. The auditory-with-bimodal overlap was computed by identifying the amount of voxels that were present in both the auditory and bimodal clusters (red dashed area) divided by the amount of auditory voxels composing the auditory clusters. Within the schematic voxels, we also included two different arrows representing the typical amount of relative responses for visual (green) and auditory (red) condition in each of the five compartments (exclusively visual, exclusively auditory, visual-bimodal common area, auditory-bimodal common area, and equally responsive bimodal voxels). Following the definition of visual, auditory, and bimodal voxels, the distribution of the type of voxels in those compartments was the following: there were exclusively visual voxels responding to the visual condition to a larger extent compared with the auditory condition but characterized by a negative activation for the auditory condition; the same behavior characterized the exclusively auditory voxels; within the visual-bimodal common region, there were voxels significantly responding more for visual than auditory conditions and both responding positively; likewise, in the auditory-bimodal common region and finally the equally responsive bimodal voxels were those that compared with the other categories were characterized by a more equal response to visual and auditory condition. Right, Same illustration as in left for the high-res dataset (in this case, visual, auditory, and bimodal clusters, voxels, or conditions are represented by light green, orange, and light blue, respectively). **b**, Results for the visual-with-bimodal (left) and auditory-with-bimodal (right) overlap for each subject for middle-res and high-res.

millimeters) were calculated; the type of the neighboring clusters was identified; and, finally, the proportion of neighboring clusters belonging to a different type was computed.

Results

The bSTC was localized for each subject ($n = 10$) via the low-res dataset, acquired at 3T. For 8 (of 10) subjects, the bSTC consisted of clusters located both in the STS (bSTS) and in the STG (bSTG) bilaterally. For one participant, we observed the bSTS and bSTG clusters unilaterally in the right hemisphere, and for one participant the bSTC was restricted to the STG (bSTG) only on the left hemisphere (Fig. 2). At middle-res and high-res, we observed that the bSTC of each subject consisted of clusters of voxels selectively responding to the visual or auditory condition (visual and auditory clusters, respectively), or to both types of stimuli (bimodal clusters) (e.g., Figs. 3b, 7b). More specifically, among the 9 subjects in whom the bSTC included clusters in the bSTS and bSTG, 7 subjects showed visual, auditory, and bimodal clusters within both regions and 2 subjects showed the three cluster types only within the bSTG. Finally, visual, auditory, and bimodal clusters were also observed within the bSTC of the participant whose bSTC corresponded to the bSTG only. We did not observe any lateralization for the 8 subjects showing a bilateral bSTC.

At middle-res, the localization of the auditory and bimodal clusters was strongly reliable ($r = 0.524$ and $r = 0.605$, respec-

tively, on average) and significant ($p < 0.05$) for all subjects. The localization of the visual clusters was also significantly reliable ($p < 0.05$, for 9 of 10 subjects) but to a lesser extent than the other two cluster types ($r = 0.299$ on average). For the high-res dataset, we observed similar results, namely, a significantly reliable localization ($p < 0.05$, for all subjects) of the auditory ($r = 0.448$ on average), bimodal ($r = 0.537$ on average), and visual ($r = 0.219$ on average) clusters.

At both middle and high resolution, the largest proportion of voxels was classified as bimodal, followed by auditory and visual voxels. The three distributions of the proportion of voxels across voxel types were significantly different ($p = 0.000$), whereas the difference of the two distributions of the proportion of voxels across resolutions was statistically indistinguishable ($p = 0.976$). Comparing each voxel type (with a nonparametric sign test) revealed that, at both high-res and middle-res, the proportion of bimodal voxels was significantly larger than the proportion of visual ($p = 0.002$ for both resolution) but not of auditory voxels ($p = 0.344$ for middle-res and $p = 0.754$ for high-res) (Fig. 3c).

The visual and auditory clusters showed a selectivity index that was almost 2 times larger at high-res compared with middle-res for all 10 subjects (1.998 and 1.619 times larger, on average, for visual and auditory selectivity index, respectively) (Fig. 4c). Nonparametric statistical analysis (sign test) showed that, at the

Figure 6. Bimodal integration index analysis. **a**, Example of activation maps (Subject 02) at the middle-res and high-res and computational procedure of the bimodal integration index (max criterion). The β weight from the GLM analysis related to the visual, auditory, and bimodal congruent condition was extracted from the bimodal clusters within the bSTC at middle-res (top) and high-res (bottom). **b**, Results for the bimodal integration index for each subject for middle-res and high-res.

group level, selectivity at high-res was larger than the selectivity at middle-res ($p = 0.002$), for both the visual and auditory index. We observed a decrease of spatial overlap between the bimodal and visual clusters (computed as the proportion of voxels being both bimodal and visual) for the high-res compared with the middle-res dataset for 9 (of 10) subjects (reduction of 54.009%, on average). For the remaining subject, no overlap was observed for either resolution. The same type of result was observed for the overlap between auditory and bimodal clusters (computed as the proportion of voxels being both bimodal and auditory), which was smaller for high-res than middle-res dataset for 9 (of 10) subjects (reduction of 48.561%, on average). For the other subject, the overlap increased with resolution (Fig. 5b). At the group level, the nonparametric statistical testing (sign test) confirmed that the overlap between the visual and the bimodal clusters was smaller for the high-res than middle-res dataset ($p = 0.004$). And in line with the single-subject analyses, the same held also for the overlap between the auditory and the bimodal clusters ($p = 0.021$). In sum, the overlap analysis revealed a larger proportion of equally responsive bimodal voxels at high-res than at middle-res. The integrative power of bimodal voxels showed larger values for high-res than middle-res dataset for 9 (of 10) subjects (for the remaining subject, there was an opposite trend; Fig. 6b), and this difference was significant at the group level ($p = 0.021$; sign test).

We finally investigated the neighborhood cluster type within the bSTC, and we observed that clusters of a given type (visual, auditory, or bimodal in the bSTC) were surrounded by a larger proportion of clusters belonging to a different type at high-res than at mid-res. Moreover, this effect occurred regardless of the distance between the target cluster and its neighbors and was significant (sign test) for distances >4.5 mm (Fig. 7b).

Discussion

In the present study, the functional properties and spatial layout of the human bSTC were mapped at middle and high spatial resolution using 7 Tesla fMRI. We observed that the bSTC (localized at low-res) was not exclusively composed by bimodal voxels, but it is spatially organized in functional clusters selectively responding to visual or auditory inputs (unimodal visual and auditory), and clusters responding to both modalities (bimodal). This result is in line with converging evidence suggesting that, to observe separate clusters of unimodal and bimodal voxels within the bSTC, it is necessary to go beyond the conventional fMRI spatial resolution (Beauchamp et al., 2004a, b; van Atteveldt et al., 2010). However, none of these previous studies was able to unravel to what extent identified unimodal and bimodal patches still reflect a mixture of different populations (Beauchamp et al., 2004b; Dahl et al., 2009).

To further address this still open question, here we investigated the topography of the bSTC first at the macroscopic level by computing the proportion of the different voxel types (i.e., unimodal visual, unimodal auditory, or bimodal) across spatial resolutions. First, we observed that letters elicited, in both resolutions, a proportion of visual voxels (6.07% for middle-res and 5.07% for high-res, on average) that was substantially smaller than the proportion of auditory voxels. This is likely caused by the high sensitivity of the bSTC to biological motion stimuli that was absent in the present study but included in the natural stimuli used in other studies (Karnath, 2001; Saygin, 2007; Saygin, 2012). A compelling empirical evidence that natural/nonlanguage stimuli and artificial/language stimuli (vowels and consonants, used here) activated the bSTC in different manners is represented by

Figure 7. Spatial organization analysis. *a*, Two examples of activation maps (Subjects 07 and 10) at middle-res and at high-res. *b*, The results from the spatial organization analysis measure the level of patchiness across resolutions. More specifically, it shows to what extent clusters belonging to a specific type (visual or auditory or bimodal) are surrounded by clusters of different type. The *y* axes represent the proportion of clusters of different types surrounding a specific target cluster at a specific distance as indicated by the *x* axes.

the proportion of visual voxels observed in the two cases. Previous high-resolution studies on multisensory integration, for example, that used natural stimuli: movie clips of tools being used (Beauchamp et al., 2004b) or monkey vocalizations (Dahl et al., 2009), showed a proportion of visual voxels within the bSTC much larger and comparable with auditory voxels (Beauchamp et al., 2004b; Dahl et al., 2009). Therefore, the difference between the two types of stimulus materials and their different effect within the bSTC confirms the flexible nature of the multisensory system (van Atteveldt et al., 2014; Murray et al., 2016) and suggests that the topographical organization of the bSTC we observed in the present study might be specific to the human cortex and related to literacy acquisition.

When we looked at the proportion of all three types of voxels, we observed that the proportion of visual, auditory, and bimodal voxels was very similar across resolutions (bimodal > auditory > visual voxels both for middle-res- and high-res datasets; Fig. 3c). The lack of differences in terms of proportion of voxels in the three categories across resolutions did not allow us to infer, at least from a “macro” point of view, whether increasing the resolution from middle-res to high-res would result in a more accurate functional and spatial organization of the bSTC.

Therefore, we investigated the functional and spatial properties of the visual, auditory, and bimodal clusters within the bSTC across high-res and middle-res datasets in more detail at the single-subject level. Comparing the functional selectivity for the three cluster types across resolution revealed that the selectivity of unimodal visual and auditory clusters was almost two times higher for the high-res than the middle-res dataset. In addition, we observed a larger proportion of equally responsive bimodal voxels (i.e., equally active for unimodal visual and unimodal auditory) for the high-res dataset. Together with the higher response selectivity observed for visual and auditory clusters for

high-res than middle-res, this result suggests that increasing the resolution of the fMRI measurement works as a filter to eliminate spurious “modality-specific” voxels. However, the overlap analysis was not able to address the important question related to bimodal integration and whether the larger proportion of equally responsive bimodal voxels at high-res was actually corresponding to a higher amount of integration of unimodal visual and auditory stimuli or it was just merely indexing the presence of a larger amount of voxels responsive to both unimodal conditions without necessarily performing interactive operations. To disentangle between these two possibilities, the integrative power of bimodal voxels was computed. We observed that such parameter increased with spatial resolution suggesting that at high-res, not only bimodal voxels consisted to a larger extent (than at middle-res) of voxels equally responding to visual and auditory but also that these voxels exhibited stronger audiovisual integration than at middle-res. The observed larger selectivity for the high versus middle spatial resolution suggests that auditory, visual, or bimodal voxels at high-res contain a larger relative amount of the corresponding (visual, auditory, or bimodal) neuronal populations than at middle-res. Thus, to characterize voxels in the bSTC substantially more accurately, acquisition of fMRI data with a resolution of at least $1.1 \times 1.1 \times 1.1$ mm³ seems recommendable.

To have a complete view of the “evolution” of the bSTC when increasing the spatial resolution, we also investigated its spatial layout. First, we observed that both the bSTC (when localized at low-res) and the three decomposed functional cluster types (identified at middle and high-res) were mainly located bilaterally (for 8 subjects and unilaterally right or left for the remaining 2). Therefore, at the “macroscopic” level, there was no differences across resolutions in terms of lateralization. Such a lack of lateralization is in line with the evidence that the speech perception system as well as the core regions involved in bimodal integration

are organized bilaterally (Beauchamp et al., 2004b; van Atteveldt et al., 2004; Beauchamp, 2005b; Humphries et al., 2005; Hickok and Poeppel, 2007; Hickok et al., 2008; Rogalsky et al., 2011).

We, then, focused on the layout of middle and high-res data to understand whether the increase of resolution, and therefore of selectivity, would be accompanied by changes in the spatial layout of the bSTC and in particular how visual, auditory, and bimodal clusters were spatially organized at the more accurate resolution (i.e., at high-res). Indeed, increasing resolution in fMRI has been proven useful to map fundamental units of sensory processing, such as cortical columns (Yacoub et al., 2007, 2008; Zimmermann et al., 2011; De Martino et al., 2015; Nasr et al., 2016), organizational units extending across cortical depth that consist of approximately hundred thousand neurons with similar response specificity (Mountcastle, 1957; Hubel and Wiesel, 1974; Fujita et al., 1992). Although we are aware that neither at $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ nor at $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ is it possible to target cortical columns within the bSTC, the results of our spatial layout analysis revealed a larger proportion of neighboring clusters of different categories for the high-res than for middle-res dataset (i.e., clusters from a specific category, e.g., visual, were more distant to different type of clusters, e.g., auditory and bimodal, at middle-res than at high-res). This indicated that the visual, auditory, and bimodal clusters at high-res were more likely intermingled with each other compared with their distribution at middle-res suggesting that the topographical organization of the categorical clusters at high-res was “more patchy” than at middle-res, which is typically expected when approaching an orderly columnar-like spatial organization.

Finally, our study underlines the importance of the single-subject fMRI analyses. The necessity of such a methodology has been largely documented both theoretically and empirically (Miller and Van Horn, 2007; Van Horn et al., 2008; Besle et al., 2013; Dubois and Adolphs, 2016; Gentile et al., 2017). Analyzing each subject individually was an effective way to preserve the spatial specificity of each subject’s dataset and allowed us to perform a detailed topographic investigation. In the present study, we indeed observed a large variability across subjects in terms of layout of unimodal and bimodal clusters; and most importantly, the single-subject analysis showed that the selectivity parameters related to the bimodal clusters (proportion of equally responsive bimodal voxels and their integrative power) largely varied across subjects (Figs. 5b, 6b). The large variability of the bimodal selectivity parameters (e.g., integrative power of bimodal voxels) highlighted here may indicate true variability, but it might also indicate that the optimal resolution to clearly separate voxel response types may be different from subject to subject. Indeed, if we assume that those parameters would reach a maximum/plateau for a given resolution, that resolution would be the optimal one for representing the modality-specific responses of the bSTC in humans. Future studies would need to consider to what extent further increasing the resolution would quantitatively improve the selectivity of unimodal and bimodal clusters for each subject independently to test at which point (in terms of resolution) a specific voxel size would represent the minimum unit of information related to a specific modality. This would enable revealing the essential layout of unimodal and bimodal regions within the bSTC to further investigate whether it is possible, for example, to resolve functional clusters for individual stimuli within the three investigated areas. In conclusion, the increase of the resolution of fMRI acquisition from $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ to $1.1 \times 1.1 \times 1.1 \text{ mm}^3$ provided new and fundamental insights reflecting a more accurate spatial representation of the human bSTC, thus further re-

ducing the gap between noninvasive multisensory research in humans and the respective animal literature.

References

- Amedi A, von Kriegstein K, van Atteveldt NM, Beauchamp MS, Naumer MJ (2005) Functional imaging of human crossmodal identification and object recognition. *Exp Brain Res* 166:559–571. [CrossRef Medline](#)
- Beauchamp MS (2005a) Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics* 3:93–113. [CrossRef Medline](#)
- Beauchamp MS (2005b) See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr Opin Neurobiol* 15:145–153. [CrossRef Medline](#)
- Beauchamp MS, Lee KE, Argall BD, Martin A (2004a) Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41:809–823. [CrossRef Medline](#)
- Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004b) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat Neurosci* 7:1190–1192. [CrossRef Medline](#)
- Besle J, Sánchez-Panchuelo RM, Bowtell R, Francis S, Schluppeck D (2013) Single-subject fMRI mapping at 7 T of the representation of fingertips in S1: a comparison of event-related and phase-encoding designs. *J Neurophysiol* 109:2293–2305. [CrossRef Medline](#)
- Calvert GA, Hansen PC, Iversen SD, Brammer MJ (2001) Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage* 14:427–438. [CrossRef Medline](#)
- Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. *J Neurosci* 29:11924–11932. [CrossRef Medline](#)
- De Martino F, Moerel M, Ugurbil K, Goebel R, Yacoub E, Formisano E (2015) Frequency preference and attention effects across cortical depths in the human primary auditory cortex. *Proc Natl Acad Sci U S A* 112:16036–16041. [CrossRef Medline](#)
- Doehrmann O, Naumer MJ (2008) Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Res* 1242:136–150. [CrossRef Medline](#)
- Dubois J, Adolphs R (2016) Building a science of individual differences from fMRI. *Trends Cogn Sci* 20:425–443. [CrossRef Medline](#)
- Fujita I, Tanaka K, Ito M, Cheng K (1992) Columns for visual features of objects in monkey inferotemporal cortex. *Nature* 360:343–346. [CrossRef Medline](#)
- Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878. [CrossRef Medline](#)
- Gentile F, Ales J, Rossion B (2017) Being BOLD: the neural dynamics of face perception. *Hum Brain Mapp* 38:120–139. [CrossRef Medline](#)
- Goebel R, van Atteveldt N (2009) Multisensory functional magnetic resonance imaging: a future perspective. *Exp Brain Res* 198:153–164. [CrossRef Medline](#)
- Heinrich MP, Papież BW, Schnabel JA, Handels H (2014) Non-parametric discrete registration with convex optimisation. In: *Biomedical image registration* (Modat S, ed), pp 51–61. New York: Springer.
- Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402. [CrossRef Medline](#)
- Hickok G, Okada K, Barr W, Pa J, Rogalsky C, Donnelly K, Barde L, Grant A (2008) Bilateral capacity for speech sound processing in auditory comprehension: evidence from Wada procedures. *Brain Lang* 107:179–184. [CrossRef Medline](#)
- Hubel DH, Wiesel TN (1974) Sequence regularity and geometry of orientation columns in monkey striate cortex. *J Comp Neurol* 158:267–293. [CrossRef Medline](#)
- Humphries C, Love T, Swinney D, Hickok G (2005) Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Hum Brain Mapp* 26:128–138. [CrossRef Medline](#)
- Karnath HO (2001) New insights into the functions of the superior temporal cortex. *Nat Rev Neurosci* 2:568–576. [CrossRef Medline](#)
- Laurienti PJ, Perrault TJ, Stanford TR, Wallace MT, Stein BE (2005) On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp Brain Res* 166:289–297. [CrossRef Medline](#)
- Miller MB, Van Horn JD (2007) Individual variability in brain activations associated with episodic retrieval: a role for large-scale databases. *Int J Psychophysiol* 63:205–213. [CrossRef Medline](#)

- Mountcastle VB (1957) Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J Neurophysiol* 20:408–434. [CrossRef Medline](#)
- Murray MM, Lewkowicz DJ, Amedi A, Wallace MT (2016) Multisensory processes: a balancing act across the lifespan. *Trends Neurosci* 39:567–579. [CrossRef Medline](#)
- Nasr S, Polimeni JR, Tootell RB (2016) Interdigitated color- and disparity-selective columns within human visual cortical areas V2 and V3. *J Neurosci* 36:1841–1857. [CrossRef Medline](#)
- Rogalsky C, Rong F, Saberi K, Hickok G (2011) Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. *J Neurosci* 31:3843–3852. [CrossRef Medline](#)
- Saygin AP (2007) Superior temporal and premotor brain areas necessary for biological motion perception. *Brain* 130:2452–2461. [CrossRef Medline](#)
- Saygin AP (2012) Sensory and motor brain areas supporting biological motion perception: neuropsychological and neuroimaging studies. In: *People watching: social, perceptual, and neurophysiological studies of body perception* (Johnson K, Shiffrar M, eds), pp 371–389. Oxford: Oxford University Press.
- Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, 'unisensory' processing. *Curr Opin Neurobiol* 15:454–458. [CrossRef Medline](#)
- Schroeder CE, Smiley J, Fu KG, McGinnis T, O'Connell MN, Hackett TA (2003) Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *Int J Psychophysiol* 50:5–17. [CrossRef Medline](#)
- Seltzer B, Cola MG, Gutierrez C, Massee M, Weldon C, Cusick CG (1996) Overlapping and nonoverlapping cortical projections to cortex of the superior temporal sulcus in the rhesus monkey: double anterograde tracer studies. *J Comp Neurol* 370:173–190. [CrossRef Medline](#)
- van Atteveldt NM, Blau VC, Blomert L, Goebel R (2010) fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neurosci* 11:11. [CrossRef Medline](#)
- van Atteveldt N, Formisano E, Goebel R, Blomert L (2004) Integration of letters and speech sounds in the human brain. *Neuron* 43:271–282. [CrossRef Medline](#)
- van Atteveldt N, Murray MM, Thut G, Schroeder CE (2014) Multisensory integration: flexible use of general operations. *Neuron* 81:1240–1253. [CrossRef Medline](#)
- Van de Moortele PF, Auerbach EJ, Olman C, Yacoub E, Ugurbil K, Moeller S (2009) T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. *Neuroimage* 46:432–446. [CrossRef Medline](#)
- Van Horn JD, Grafton ST, Miller MB (2008) Individual variability in brain activity: a nuisance or an opportunity? *Brain Imaging Behav* 2:327–334. [CrossRef Medline](#)
- Wallace MT, Ramachandran R, Stein BE (2004) A revised view of sensory cortical parcellation. *Proc Natl Acad Sci U S A* 101:2167–2172. [CrossRef Medline](#)
- Werner S, Noppeney U (2010) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J Neurosci* 30:2662–2675. [CrossRef Medline](#)
- Yacoub E, Shmuel A, Logothetis N, Ugurbil K (2007) Robust detection of ocular dominance columns in humans using Hahn Spin Echo BOLD functional MRI at 7 Tesla. *Neuroimage* 37:1161–1177. [CrossRef Medline](#)
- Yacoub E, Harel N, Ugurbil K (2008) High-field fMRI unveils orientation columns in humans. *Proc Natl Acad Sci U S A* 105:10607–10612. [CrossRef Medline](#)
- Zimmermann J, Goebel R, De Martino F, van de Moortele PF, Feinberg D, Adriany G, Chaimow D, Shmuel A, Ugurbil K, Yacoub E (2011) Mapping the organization of axis of motion selective features in human area MT using high-field fMRI. *PLoS One* 6:e28716. [CrossRef Medline](#)