Category Training Induces Cross-modal Object Representations in the Adult Human Brain

Marieke van der Linden[1], Miranda van Turennout[1], and Guillén Fernández[1,2]

Abstract

■ The formation of cross-modal object representations was investigated using a novel paradigm that was previously successful in establishing unimodal visual category learning in monkeys and humans. The stimulus set consisted of six categories of bird shapes and sounds that were morphed to create different exemplars of each category. Subjects learned new cross-modal bird categories using a one-back task. Over time, the subjects became faster and more accurate in categorizing the birds. After 3 days of training, subjects were scanned while passively viewing and listening to trained and novel bird types. Stimulus blocks consisted of bird sounds only, bird pictures only, matching pictures and sounds (cross-modal congruent), and mismatching pictures and sounds (cross-modal incongruent). fMRI data showed unimodal and cross-modal training effects in the right fusiform gyrus. In addition, the left STS showed cross-modal training effects in the absence of unimodal training effects. Importantly, for both the right fusiform gyrus and the left STS, the newly formed cross-modal representation was specific for the trained categories. Learning did not generalize to incongruent combinations of learned sounds and shapes; their response did not differ from the response to novel cross-modal bird types. Moreover, responses were larger for congruent than for incongruent cross-modal bird types in the right fusiform gyrus and STS, providing further evidence that categorization training induced the formation of meaningful cross-modal object representations. ■

INTRODUCTION

We can rapidly discriminate a pigeon from a chicken, not only by looking at it but also by listening to it. The image and the sound of an object are tightly linked and provide clues for its categorization. In this study, we investigated the formation of cross-modal object representations in the human brain resulting from cross-modal category learning.

Increased visual experience with object categories has been linked to neuronal changes in category-selective areas in occipito-temporal cortex. Specifically, learning to discriminate objects from a novel category modulates activity in the right middle fusiform gyrus (van der Linden, Murre, & van Turennout, 2008; Weisberg, van Turennout, & Martin, 2007; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999) and lateral occipital gyrus (Op de Beeck, Baker, DiCarlo, & Kanwisher, 2006). Activity in occipito-temporal cortex has also been found to be selectively enhanced for objects from a category with which subjects have extensive experience, such as birds and cars (Xu, 2005; Gauthier, Skudlarski, Gore, & Anderson, 2000), or Lepidoptera (Rhodes, Byatt, Michie, & Puce, 2004). In a previous study (van der Linden, van Turennout, & Indefrey, 2010), we found the STS to be involved in the formation of associations between perceptually different exemplars within a category.

For the formation of cross-modal object representations, the role of association also seems crucial. Early in life, we need to learn which shapes and sounds of objects belong together. Indeed, the superior temporal has also been found to be involved in associating familiar sounds and shapes to facilitate cross-modal object representations. Common cross-modal objects, such as animals and tools, elicited enhanced responsiveness of posterior STS compared with unimodal stimuli (Beauchamp, Lee, Argall, & Martin, 2004). Cross-modal categories that are acquired later in life, such as letters and speech sounds, were also found to activate the superior temporal gyrus and sulcus (van Atteveldt, Formisano, Blomert, & Goebel, 2007; van Atteveldt, Formisano, Goebel, & Blomert, 2004, 2007). Recently, it became clear that familiar cross-modal objects activated the STS but not novel artificial cross-modal objects, indicating that audiovisual integration is influenced by familiarity (Hein et al., 2007). Therefore, it seems likely that cross-modal representations, such as found in the STS, can be shaped as a result of experience with cross-modal objects. This has been tested by Naumer et al. (2009). After training subjects to associate eight nonsense objects with sounds, they found more activity in frontal, parietal, and cingulate areas of the brain compared with pretraining.

[1]Radboud University Nijmegen, The Netherlands, [2]Radboud University Nijmegen Medical Centre, The Netherlands

However, showing that an area responds more to cross-modal trained than to cross-modal pretraining or novel stimuli does not automatically mean that this region is also involved in a meaningful cross-modal representation. It could simply mean that mere exposure alone is enough to induce plasticity in these areas. If cross-modal integration is successful and the representation is meaningful, the brain regions involved should show a dissociation between congruent (sound and shape match, meaningful) and incongruent (sound and shape do not match, meaningless) cross-modal stimuli. Therefore, congruency effects are usually investigated to make inferences about cross-modal integration and representations (Taylor, Moss, Stamatakis, & Tyler, 2006; van Atteveldt et al., 2004; Calvert, Campbell, & Brammer, 2000). Naumer et al. (2009) reported congruency effects for newly learned cross-modal objects in inferior frontal cortex and posterior middle temporal gyrus. The interplay of learned associations between vision and sound has been subject of a number of fMRI studies (for a review, see Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Calvert & Lewis, 2004). However, there has so far been no direct investigation of changes that occur in the brain as a result of acquiring entirely new cross-modal object categories.

In the present study, we used a paradigm that has successfully been applied to visual object category learning in human subjects (van der Linden et al., 2008, 2010; Jiang et al., 2007) and monkeys (Freedman, Riesenhuber, Poggio, & Miller, 2001, 2003). Subjects will learn new cross-modal categories of artificial birds (see Figure 1). The novelty in the present study is that we morphed the birds not only in the visual modality but also in the auditory modality. The boundary between the categories is expressed by information from both auditory and visual modalities. Our categories are perceptual based: Birds that have the same shape and sound belong in the same category. We expect that at the end of training cross-modal object representations have been formed. Training-induced improvements in unimodal object recognition usually result in increased cortical responses to trained objects compared with responses to novel objects (van der Linden et al., 2008; Weisberg et al., 2007; Moore, Cohen, & Ranganath, 2006; Op de Beeck et al., 2006; Gauthier et al., 1999). We expect regions that are involved in training-dependent cross-modal representations to show more activity for trained cross-modal congruent birds than for novel cross-modal birds. However, some training-related decreases in activation as a result of repeated stimulus exposure can also occur (Grill-Spector, Henson, & Martin, 2006). Regions showing training-related increases in activity should enclose at least the right fusiform gyrus and the STS. Importantly, if these regions are involved in a meaningful representation of cross-modal objects, they should show no training effect for incongruent stimuli. Moreover, these areas should show a congruency effect, dissociating between congruent and incongruent cross-modal bird stimuli (Doehrmann & Naumer, 2008). In addition, the inferior frontal gyrus will likely show the opposite pattern of response. The inferior frontal gyrus' responses are modulated by the meaningfulness (or semantics) of cross-modal stimuli (Doehrmann & Naumer, 2008) and usually shows a higher response to incongruent stimuli (Hein et al., 2007; Olivetti Belardinelli et al., 2004).

METHODS

Subjects

Sixteen healthy participants (5 men, mean age = 21.6 years, range = 18–26 years) participated in the experiment. All subjects had normal or corrected-to-normal vision and no hearing problems. Subjects were paid for their participation. All subjects gave written informed consent.

Stimuli

Shapes

The same stimuli were used as in van der Linden et al. (2008, 2010). The stimuli consisted of pictures of computer-generated birds that were constructed in a 3-D model manipulation program (Poser 4; Curious Labs, Santa Cruz, CA). First, six prototype birds were constructed from a base-bird (Songbird Remix; Daz3d, Draper, UT; see Figure 1A). Parts of the bird that were manipulated included its trunk, tail, beak, head shape, cheeks, brow, and eye position. Next, each of the six birds was morphed with all other birds. The category boundary was set at 50% (Figure 1C). As a result, stimuli that were close to but on opposite sides of the category boundary were visually similar but belonged to different categories. Morphing happened smoothly between corresponding points on the birds. Each bird was colorless, rendered under the same lighting and camera settings, and exported as an image. Images had identical color, shading, and scale. The images measured 300 × 300 pixels in the training sessions and were slightly reduced in size (250 × 250 pixels) in the scanning sessions.

Sounds

For the auditory stimuli, six sound fragments were taken from real bird calls, see Supplementary Table 1. These sound fragments were converted to wave files with a sampling rate of 44 kHz and multiplied with a Gaussian (see Figure 1B). The length and loudness of the sounds was matched; each sound measured 500 msec, and the loudness was set to 80 dB for all wave files. Finally, the wave files were morphed with each other in the same ratios as the visual stimuli using the formula: morphed sound A:B = (morph ratio × amplitude sound A) + ((1 − morph ratio) × amplitude sound B) (see Figure 1C). All described manipulations were done using the Praat software (http://www.praat.org).

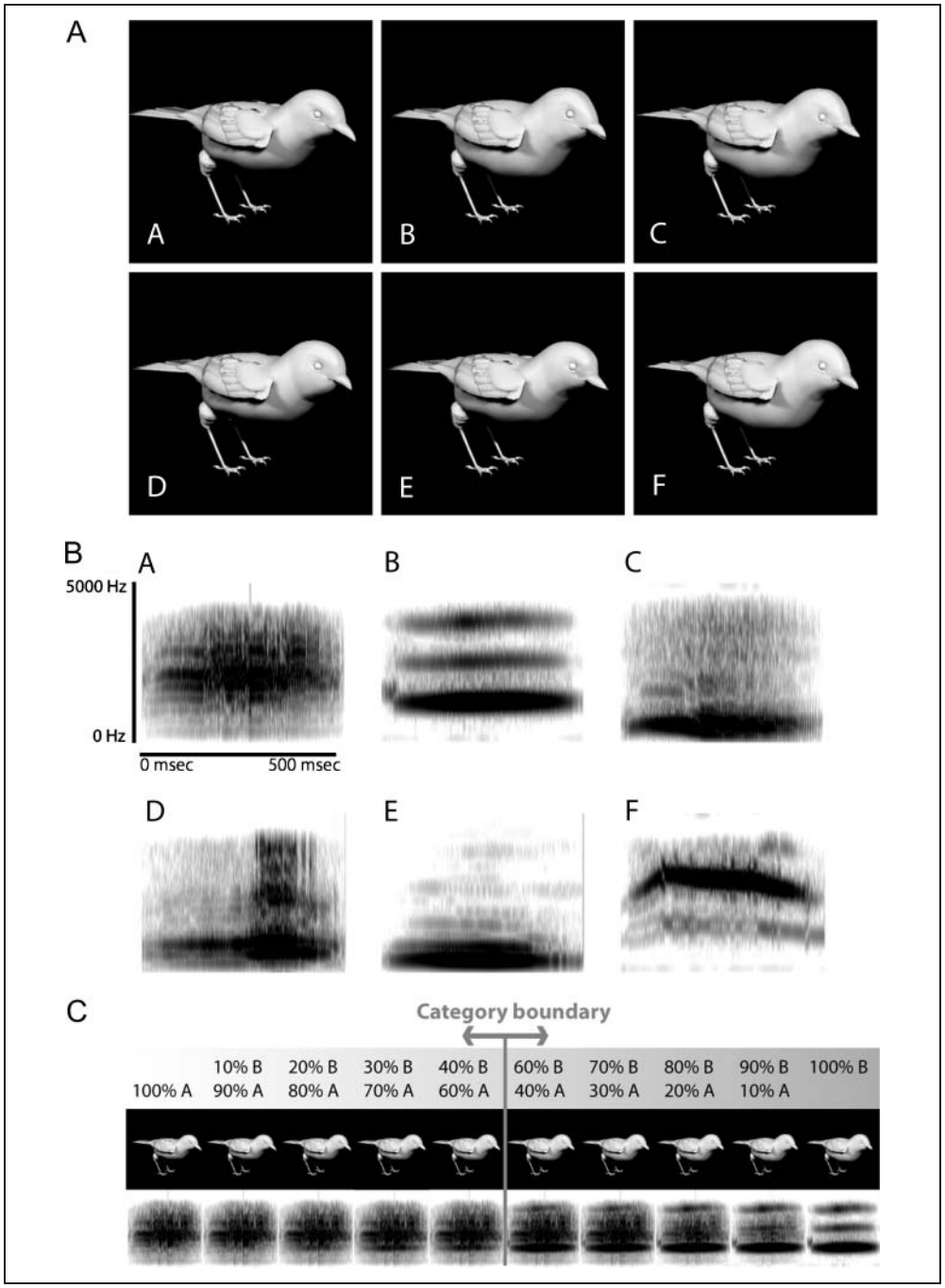

Figure 1. Construction of the stimulus set. (A) Pictures of nonexisting but plausible bird shapes were constructed in a 3-D model manipulation program. From a base bird, we derived six colorless prototype birds (A, B, C, D, E, and F) that differed in trunk, tail, beak, head shape, cheeks, brow, and eye position. Each bird was rendered under the same lighting and camera settings to make sure that shading and scale were identical for all birds. (B) Spectrogram of the bird sounds corresponding to the bird shapes. (C) Exemplars and their corresponding sounds were created by systematically morphing each of the six prototype birds with all other birds. Shown is an example of morphing the shapes (top) and sounds (bottom) of bird type A with bird type B at morph ratios of 90:10, 80:20, 70:30, and 60:40. The category boundary was set at 50:50.

Procedure

Bird shapes and sounds were paired to create cross-modal bird stimuli. The pairing of sounds and shapes was arbitrary. The morph ratio between shape and sound always corresponded (i.e., 70% bird type A morphed with 30% bird type B would also have the sound of 70% bird type A morphed with 30% bird type B). Three bird types were assigned to be trained, and three bird types acted as novel controls during scanning. The bird types constituting the trained and novel conditions were counterbalanced across subjects. Birds were only morphed with each other within a condition, so if the trained bird types were A, B, and C, the exemplars would consist of morphs of 55%, 65%, 70%, 80%, and 95% of A:B, B:A, A:C, C:A, B:C, and C:B.

Training

Training included three sessions on separate days, each of which lasted approximately 1 1/2 hours. During a training session, subjects sat comfortably in a soundproof cabin in front of a 19-in. computer screen to view the bird shapes. Subjects wore a headphone to listen simultaneously to the bird sounds. During training, they performed a one-back task on a series of cross-modal bird stimuli (Figure 2A), in

which they indicated with the index and middle finger of their right hand whether two consecutive birds were the same bird type or not. Subjects received feedback to their responses consisting of a printed text centered on the screen in colored Arial font in size 16 (green: "right"; red: "wrong"; and yellow: "too late"). During one block of training, two cross-modal bird types would be presented. There were 10 exemplars (each bird type was morphed at five morph levels with the other two bird types) for each of the three trained bird types. Each exemplar was presented 45 times per training session. The proportion of birds from the same and different categories was fifty-fifty. In each trial, stimuli were presented for 1000 msec after which a response could be given during 2250 msec. Feedback was presented for 250 msec. Stimuli onset asynchrony was 4000 msec. A training session consisted of nine blocks of 100 trials. Each block of 100 trials was followed by a small self-paced pause after which a subject could continue the experiment by pressing a button. After five blocks of training, the subjects had a longer break during which they left the soundproof cabin and drank coffee or tea.

fMRI Scanning Session

Subjects participated in an fMRI scanning session one day after training. During scanning, subjects were presented with trained and novel bird stimuli in blocks (Figure 2B). Stimulus blocks consisted of bird sounds only, bird shapes only, matching pictures and sounds (cross-modal congruent), and mismatching pictures and sounds (cross-modal incongruent). Bird exemplars consisted of morph levels that were different from the morph levels that the subjects trained with to avoid simple repetition effects. Morph levels were 60%, 75%, and 90% and were presented pseudorandom within the blocks. Each block contained nine bird stimuli at three morph levels. Each image was presented for 1 sec and each sound for 500 msec (with a simultaneously presented fixation cross of 1 sec), with a mean interstimulus interval of 2 sec (varying random between

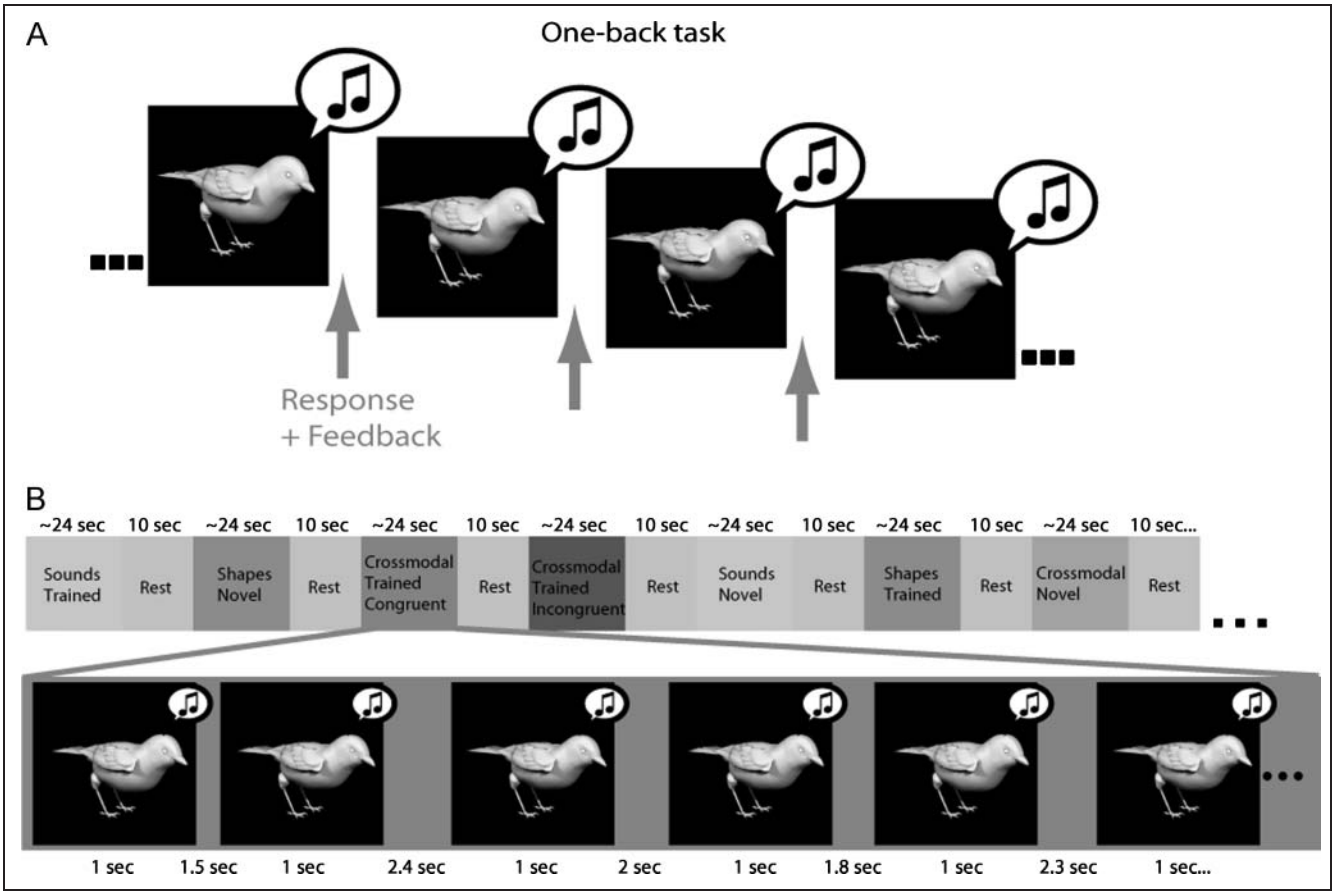

Figure 2. Training and fMRI paradigms. (A) During the training sessions, participants were presented with a series of cross-modal bird exemplars. They performed a one-back task in which they indicated whether two consecutive birds were the same type or not. Category learning was established by providing corrective feedback after each trial. (B) In the posttraining fMRI scanning session, the bird types were presented in blocks of 10 exemplars at mixed morph ratios of 60:40, 75:25, and 90:10. Stimulus blocks consisted of bird sounds only, bird pictures only, matching pictures and sounds (cross-modal congruent), and mismatching pictures and sounds (cross-modal incongruent). Blocks consisted of either trained or novel birds. Each bird was presented for 1 sec with a mean interstimulus interval of 2 sec. Experimental blocks alternated with rest periods of 10 sec. Subjects were instructed to view and listen to the birds attentively.

1500 and 2500 msec). Experimental blocks lasted 25 sec and alternated with rest periods of 10 sec for sampling the baseline. Blocks were presented 10 times per condition in pseudorandom order. For each morph level, there were 30 trials. Total scan time was 47 min. Subjects were instructed to view and to listen attentively to the birds. We were interested in investigating the automatic activation of cortical object representations; therefore, we have chosen a passive paradigm to minimize task-related activation. Task instructions have an effect on the automatic integration of sound and percept (de Gelder & Bertelson, 2003) and can even overrule it (van Atteveldt, Formisano, Goebel, et al., 2007). A passive task is widely used to investigate automatic processing of unimodal and cross-modal stimuli (Hein et al., 2007; Belardinelli et al., 2004; van Atteveldt et al., 2004; Calvert et al., 2000), also for studies that combined the scanning session with a learning phase (Naumer et al., 2009).

During scanning, subjects' heads were fixated with cushions attached to the head coil. An LCD projector projected mirror-reversed stimuli on a screen at the end of the bore, which the subject was able to see through a mirror attached to the head coil. Auditory stimuli were presented using headphones (Commander XG; Resonance Technology Inc., Northridge, CA) with padding that also attenuated gradient noise. Before starting the experiment, the sound level was determined by exposing the subject to the gradient noise accompanying epi-scanning and presenting the bird sounds simultaneous. The subjects indicated at which sound level they could clearly hear the bird sounds. This sound level was then used throughout the experiment.

Imaging Parameters

For each subject, 1,300 whole-brain images (EPI, 32 slices, 3 mm thick with 10% gap, repetition time = 2170 msec, voxel size = 3 × 3 × 3 mm, echo time = 30, flip angle = 75°, field of view = 19.2 cm, matrix size = 64 × 64) were acquired on a 3-T whole-body MR scanner (Magnetom TIM TRIO; Siemens Medical Systems, Erlangen, Germany). In addition, a high-resolution structural T1-weighted 3-D magnetization-prepared rapid acquisition gradient-echo sequence image was obtained after the functional scan (192 slices, voxel size = 1 × 1 × 1 mm).

Behavioral Data Analysis

Mean response times for the correct trials and the mean proportion of correct trials were computed for each subject. These dependent variables were submitted to a Training Session × Morph Level MANOVA with repeated measures. Training session consisted of three levels (first, second, and third training session) and morph level consisted of five levels (55%, 65%, 70%, 80%, and 95%). Differences between training sessions were explored with MANOVA with two levels for training session and five levels for morph level. We investigated the differences within training sessions by examining the effect of block on accuracy with a Session × Block × Morph Level MANOVA. Analyses of separate sessions were performed using a Block × Morph Level MANOVA. Block consisted of nine levels (there were nine blocks of training per session).

fMRI Data Analysis

Imaging data analysis was done using BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). The first three volumes were discarded to allow for T1 signal equilibrium. The following preprocessing steps were performed: slice scan time correction (using sinc interpolation), linear trend removal, temporal high-pass filtering to remove low-frequency nonlinear drifts of three or fewer cycles per time course, and 3-D motion correction to detect and to correct for small head movements by spatial alignment of all volumes to the first volume by rigid body transformations. Estimated translation and rotation parameters were inspected and never exceeded 3 mm. Coregistration of functional and 3-D structural measurements was computed by relating functional images to the structural scan, which yielded a 4-D functional data set. Structural 3-D and functional 4-D data sets were transformed into Talairach space (Talairach & Tournoux, 1988).

The inhomogeneity-corrected structural scans were used for individual subjects' cortex reconstruction (Kriegeskorte & Goebel, 2001). For each individual subject, the gray and the white matter were segmented. The border between white and gray matter was used to produce a surface reconstruction of each hemisphere. To improve the spatial correspondence between subjects' brains beyond Talairach space matching, the reconstructed hemispheres were aligned using curvature information reflecting the gyral/sulcal folding pattern. Folded cortical representations of each subject and hemisphere were morphed into a spherical representation. These spherical representations were aligned to one another using an algorithm accounting for an optimal fit of the main gyrification with minimal distortion between the individual cortices. Alignment of major gyri and sulci was achieved reliably using this method. Cortex-based intersubject alignment enabled us to align the time courses for multisubject general linear model (GLM) data analysis. Group-averaged functional data were then projected on inflated representations of the left and right cerebral hemispheres of a single subject.

Cortex-based statistical analysis was performed using multiple linear regression. For every cortical surface vertex, the time course was regressed on a set of predictors representing our eight experimental conditions. Regressors of interest were modeled using a gamma function ($\tau = 2.5$ sec, $\delta = 1.5$) convolved with the blocks of experimental conditions (Boynton, Engel, Glover, & Heeger, 1996). Because for novel birds there existed no representation of congruent or incongruent combinations, these were collapsed. In addition, six regressors of no interest representing the motion parameters were included in the

model. Multiple regression, fixed effects, was performed using the GLM. Unimodal and cross-modal activations were investigated with the following contrasts: first, unimodal activation for sounds presented in isolation: Sounds (Trained + Novel) > Rest; second, unimodal activation for shapes presented in isolation: Shapes (Trained + Novel) > Rest; and third, cross-modal activations: Cross-modal (Congruent Trained + Incongruent Trained + Novel) > Rest. Cross-modal training effects were investigated with the contrast Cross-modal Congruent Trained > Cross-modal Novel. Congruency effects were investigated with the contrast Cross-modal Trained Congruent > Cross-modal Trained Incongruent. The effect of morph level for trained birds was investigated with the contrast 90% morph level (Trained Sounds + Trained Shapes + Cross-modal Congruent + Cross-modal Incongruent) > 60% morph level (Trained Sounds + Trained Shapes + Cross-modal Congruent + Cross-modal Incongruent) and for novel birds with the contrast 90% morph level (Novel Sounds + Novel Shapes + Cross-modal Novel) > 60% morph level (Novel Sounds + Novel Shapes + Cross-modal Novel).

To correct for multiple comparisons, the false discovery rate (FDR) controlling procedure was applied on the resulting p values for all voxels. The value of q specifying the maximum FDR tolerated on average was set to .001 for overall cross-modal and unimodal activations and to .01 for cross-modal training and congruency effects. With a q value of .01, a single-voxel threshold is chosen by the FDR procedure, which ensures that from all voxels shown as active, only 1% or less are false-positives (Genovese, Lazar, & Nichols, 2002; Benjamini & Hochberg, 1995). In addition, a cluster threshold of 25 mm^3 was applied.

Significantly activated clusters were further explored with an ROI analysis in which we tested for unimodal and cross-modal training effects and for a cross-modal congruency effect. The subject-averaged responses for each condition averaged over all significantly activated voxels in a region were submitted to two-tailed paired t tests (df = 15). The tests for unimodal training effects were trained shapes versus novel shapes and trained sounds versus novel sounds. The test for cross-modal training effects was cross-modal congruent trained versus cross-modal novel and cross-modal incongruent trained versus novel cross-modal. Congruency effects were tested by testing for cross-modal congruent trained versus cross-modal incongruent trained. For these tests, an alpha level of .05 was used.

We used a psychophysiological interactions (PPI) analysis (Friston et al., 1997) to search for regions that were connected to the left STS as a result of cross-modal training. The superior temporal seed region was defined as the area that responded more to cross-modal congruent than to cross-modal novel birds ($p < .05$, FDR corrected). We used the time course from the left STS ROI as our seed region and convolved this with the vector of our contrast of interest (Cross-modal Congruent > Cross-modal Novel). This PPI regressor was then entered into a GLM together with the time course of the seed region and the vector that represented the contrast itself. The GLM estimated those voxels where there was a significant change in connectivity between cross-modal congruent and cross-modal novel birds. The threshold of this analysis was at $p < .05$ (FDR corrected).

RESULTS

Training

Analysis of the behavioral training data showed that participants became skilled in categorizing the bird exemplars (see Figure 3). The proportion of correct responses increased as training progressed over time, $F(2,14) = 72.05$, $p < .001$. Performance increased significantly from the first to the second training session, $F(1, 15) = 48.10$, $p < .001$, and from the second to the third training session, $F(1, 15) = 30.43$, $p < .001$. Within the first training session, accuracy increased from the first to the last block, $F(8, 8) = 8.67$, $p < .005$. In the second training session, there was a trend toward increased performance over blocks, $F(8, 8) = 3.10$, $p < .07$, but not in the third training session, $F(8, 8) = 1.21$, $p = ns$. In the third session, performance did not even differ between the first and the last block, $F(1, 15) = 2.05$, $p = ns$ (see Figure 3C). Although it was not our goal to have learning saturation, we did observe that training accuracy did not further increase during the last training session. We also found an effect of morph level. Responses were least accurate for birds closest to the category boundary, $F(4, 12) = 398.78$, $p < .001$. The effect of morph level was present in all training sessions: first session, $F(4, 12) = 138.22$, $p < .001$; second session, $F(4, 12) = 197.93$, $p < .001$; and third session, $F(4, 12) = 113.07$, $p < .001$.

We also found that our subjects became faster over training sessions, $F(2, 14) = 7.44$, $p < .01$. Subjects were significantly faster in the second training session than in the first training session, $F(1, 15) = 13.70$, $p < .005$, and faster in the third training session than in the second training session, $F(1, 15) = 7.06$, $p < .05$. Subjects responded faster to birds closer to the prototype, $F(4, 12) = 9.99$, $p < .005$.

fMRI Results

Subjects trained for 3 days with the cross-modal bird categories. After training, the subjects were scanned. Subjects were presented with unimodal and cross-modal bird stimuli presented in blocks during scanning. These were different exemplars than the subjects trained with. For unimodal bird types, the stimuli consisted of either bird shapes or bird sounds presented in isolation. The cross-modal bird types consisted of trained congruent (sound and shape are matching), incongruent (sound and shape do not match), and novel cross-modal bird types. Compared with unimodal bird stimuli, the cross-modal bird types activated bilateral inferior and middle frontal gyri, supramarginal gyrus, middle and superior temporal gyri,

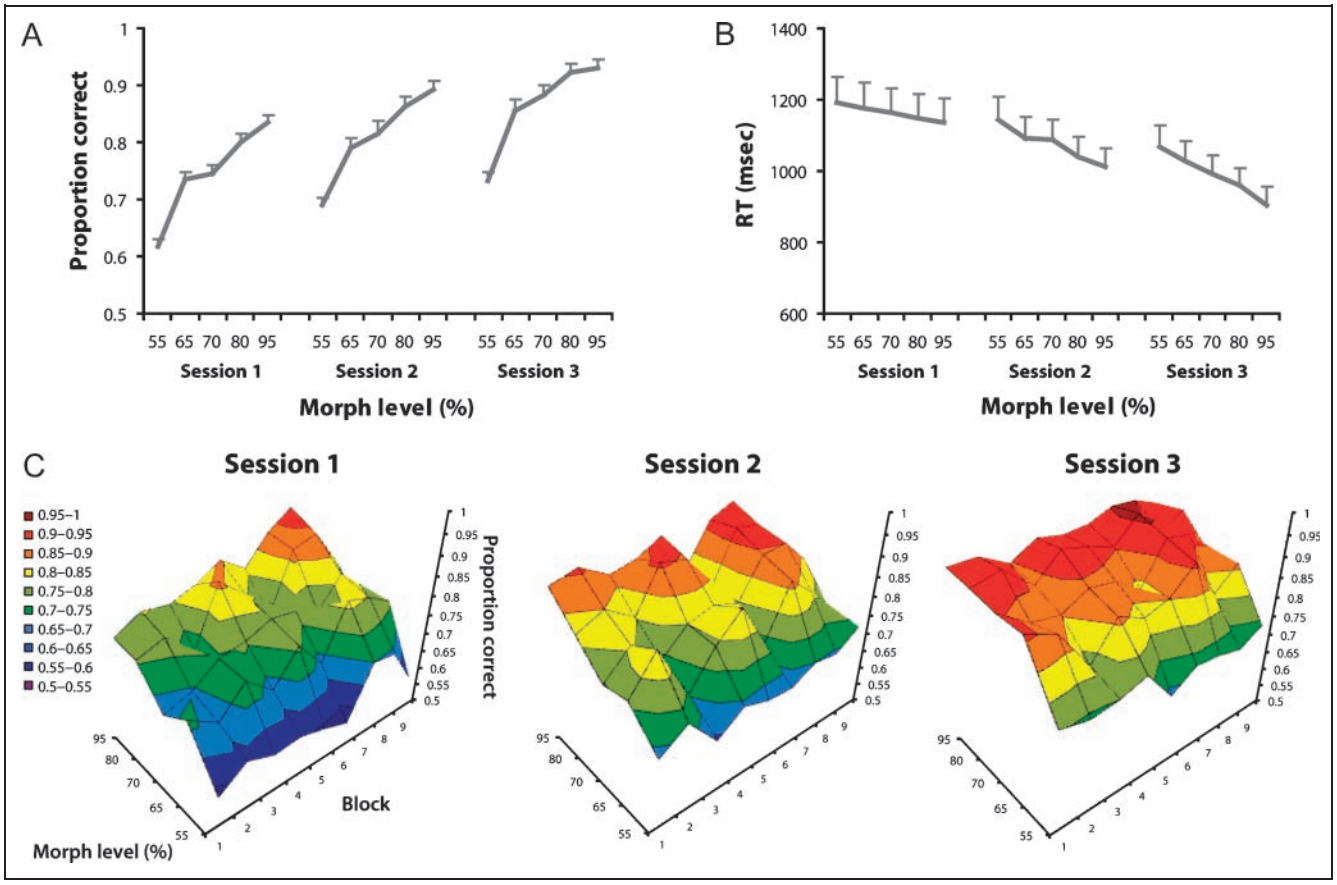

Figure 3. Training results. Mean proportion of correct responses (A) and mean response latencies (B) to the one-back task, as a function of morph level, plotted for each of the three training sessions. Error bars represent *SEM*. (C) Accuracy (proportion of correct responses) plotted as function of morph level and blocks for all three training sessions.

lateral occipital gyrus, and right STS (see Supplementary Figure 1). We also investigated which areas are responsive to both modalities, that is, to shapes and sounds presented in isolation. These areas overlap with areas that prefer cross-modal over unimodal stimuli (bilateral inferior and middle frontal gyri, bilateral lateral occipital gyri), but they exclude the superior and middle temporal gyri and include both left and right superior temporal sulci.

Cross-modal Training Effects

To investigate cross-modal training effects, we compared the responses with congruent cross-modal birds from the trained categories with responses to novel cross-modal bird stimuli at $p < .01$ (FDR corrected; see Figure 4A). The regions that were obtained from this analysis were further explored with two-tailed paired t tests ($df = 15$). We tested whether the regions showing a cross-modal training effect for congruent bird types also showed a training effect for incongruent bird types. In addition, we tested whether these regions showed a training effect for shapes and sounds presented in isolation (see Table 1).

As expected, the left STS showed a cross-modal training effect for congruent cross-modal bird types (see Figure 4B). Other regions that showed a significant cross-modal training effect were the right fusiform gyrus, the left superior temporal gyrus, the bilateral supramarginal gyrus, the left inferior frontal gyrus, the bilateral precentral gyrus, the left anterior cingulate gyrus and sulcus, the bilateral superior frontal gyrus, the bilateral insula, and the left parieto-occipital sulcus. In addition to these increases, we found that the right middle temporal gyrus showed a training-related decrease in activity.

Training effects do not necessarily indicate that the areas that showed such an effect are truly representing the newly learned categories. Mere exposure might also contribute to finding "simple" training effect. If areas are part of a meaningful cross-modal representation, the training effect should not generalize to incongruent but trained bird sound combinations. Therefore, we tested whether any of these regions showed a general training effect (see Table 1). We found that none of the regions showed a general cross-modal training effect. The response to incongruent trained bird types was never larger than the response to novel bird types.

Unimodal Training Effects

Next to these cross-modal training effects, we tested the areas that showed a cross-modal training effect for unimodal training effects. The only region that showed unimodal

training effects was the right fusiform gyrus (see Figure 4B and Table 1). The right fusiform responded more to trained bird shapes compared with novel bird shapes. In addition, responses were larger for trained bird sounds than for novel bird sounds.

To further investigate the spatial distribution of the different training effects in the right fusiform gyrus, we overlaid separate unimodal and cross-modal contrasts in the right fusiform gyrus (see Figure 5). As can be seen in Figure 5, the areas are overlapping each other largely. The training effect for the shapes extends the largest region ($x = 37, y = -23, z = -20$; 1,306 mm^3). The auditory training effect is smaller and located slightly more posterior ($x = 38, y = -35, z = -19$; 279 mm^3). The cross-modal training effect and the congruency effect are closest together in location and size (cross-modal training effect: $x = 38, y = -24, z = -18$, 279 mm^3; congruency effect: $x = 38, y = -27, z = -18$, 250 mm^3).

Congruency Effects

We found that most areas that showed a training effect also responded significantly more to the congruent bird types than to the incongruent ones. The fusiform gyrus showed higher responses for congruent cross-modal birds than for incongruent cross-modal birds (Table 1). Next to the right fusiform gyrus, the left STS also showed a congruency effect (see Figure 4B and Table 1). Other regions that showed a congruency effect were the left superior temporal gyrus, the right supramarginal gyrus, the bilateral precentral gyrus, the left cingulate sulcus, the left parieto-occipital sulcus, and the bilateral insula (see Table 1). Again, the right middle temporal gyrus, which also showed a training-related decrease, showed the reverse effect and responded more to incongruent than to congruent cross-modal birds (see Table 1).

The ROI analysis of congruency effects might be in part biased because of the contrast we used to test for training effects. The contrast already contained cross-modal congruent birds and therefore will yield those areas that have a preference for cross-modal congruent birds. If one then compares responses to congruent cross-modal birds with responses to birds from another condition within these areas, it is more likely to obtain a difference. Therefore, we also directly tested for congruency effects in the brain by contrasting congruent cross-modal birds with

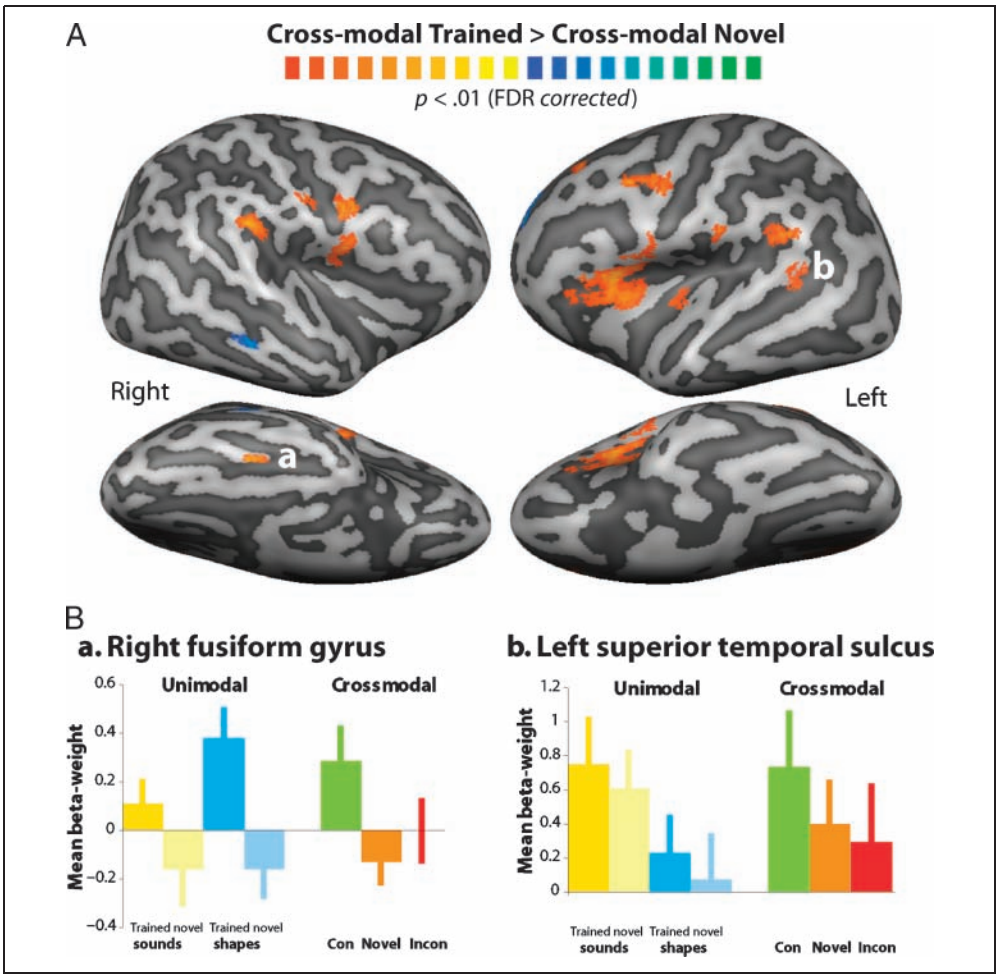

Figure 4. Cross-modal training effects. (A) Group-averaged activation maps from the posttraining scanning overlaid on lateral (top) and ventral (bottom) views of Talairach-normalized inflated hemispheres. In orange tones, regions that showed more activity for trained congruent cross-modal bird types compared with novel cross-modal bird types at $p < .01$ (FDR corrected). In blue, brain regions showing less activity following presentation of trained congruent cross-modal bird types compared with novel cross-modal bird types. (B) Voxel-averaged plots of the mean beta-weights in the left STS (A; Talairach coordinates: $x = -48, y = -51, z = 12$) and right fusiform gyrus (B; $x = 38, y = -29, z = -19$). Shown are the averaged responses for unimodal bird stimuli (sounds in yellow and shapes in blue) and cross-modal bird stimuli. For unimodal stimuli divided in trained (dark colors) and novel bird types (light colors). For cross-modal divided in trained congruent (*con* in green) and trained incongruent (*incon* in red) and cross-modal novel bird types (orange). Error bars represent *SEM*.

Table 1. Regions Showing a Cross-modal Training Effect

Area	*x*	*y*	*z*	*mm*^{3}	*AT > AN*	*VT > VN*	*CCT > CN*	*CIT > CN*	*CCT > CIT*
Training-related Increases									
R fusiform G	38	−29	−19	245	2.19*	3.13**	3.39***	0.99	2.61*
L superior temporal S	−48	−51	12	415	0.75	0.84	3.31**	−0.56	3.13**
L supramarginal G	−52	−48	30	862	0.72	0.50	3.13**	0.49	1.88
R supramarginal G	52	−36	34	787	−1.25	−1.20	2.83*	−0.93	2.49*
L superior temporal G	−35	−23	6	293	0.18	−1.74	2.40*	−0.93	3.64***
L inferior frontal G	−30	26	13	219	1.12	0.32	1.92	0.92	0.93
	−50	0	10	519	0.68	−0.34	2.60*	−2.55*	4.95****
L postcentral G	−60	−24	21	387	1.45	−0.44	2.85*	−0.91	3.80***
R postcentral G	52	−12	34	280	−0.11	−0.35	2.67*	−0.35	3.26**
L precentral G	−47	−9	42	1147	1.88	−0.22	2.51*	1.11	1.37
R precentral G	53	−6	30	757	1.24	−0.55	3.57***	0.15	2.63*
L anterior cingulate G	−2	24	8	420	1.60	−2.36*	2.54*	0.12	2.47*
L posterior cingulate G	−7	−45	32	618	−0.50	−1.03	2.08	1.54	1.21
L cingulate S	−15	−37	38	60	0.26	−1.64	2.75*	0.14	2.17*
L superior frontal G	−7	2	53	791	0.59	0.05	2.40*	0.00	1.73
R superior frontal G	8	−1	48	594	0.72	−0.10	2.53*	0.49	1.41
L parieto-occipital S	−10	−64	12	1645	−0.95	−0.70	2.87*	0.14	4.46****
L insula	−32	6	16	2417	0.91	−0.61	2.57*	0.06	2.83*
R Insula	50	−4	14	1060	0.59	−1.15	2.76*	−1.14	4.25****
Training-related Decrease									
R middle temporal G	60	−35	−6	228	0	−0.58	−2.77*	−0.26	−2.32*

Mean Talairach coordinates, volume in cubic millimeters, and averaged *t* values for regions showing a cross-modal training effect at $p < .01$, FDR corrected. In addition, we present *t* values obtained from paired *t* tests ($df = 15$) on the subject-averaged beta-weights. We tested for both unimodal training effects, auditory training effect: Trained Sounds > Novel Sounds (AT > AN) and visual training effect: Trained Shapes > Novel Shapes (VT > VN). Next, we tested for cross-modal training effects: Cross-modal Congruent Trained > Cross-modal Novel (CCT > CN) and Cross-modal Incongruent Trained > Cross-modal Novel (CIT > CN). And finally, for the congruency effect: Cross-modal Congruent Trained > Cross-modal Incongruent Trained (CCT > CIT).

L = left; R =right; G = gyrus; S = Sulcus.

*$p < .05$.

**$p < .01$.

***$p < .005$.

****$p < .001$.

cross-modal incongruent birds at $p < .01$ (FDR corrected; see Figure 6 and Table 2). This analysis confirmed the congruency effects that were obtained in the ROI analysis. In addition, some areas, including the right inferior frontal gyrus, were revealed that preferred incongruent stimuli above congruent stimuli. Within the areas showing congruency effects, we also tested for cross-modal training effects (Table 2). In addition, testing the areas that showed a congruency effect for a training effect also confirmed the previous analysis of the cross-modal training effects, being the right fusiform gyrus, the left STS and gyrus, the bilateral insula, the left inferior frontal gyrus, the right supramarginal gyrus, and the right precentral gyrus. The right inferior frontal gyrus showed the reverse pattern and preferred novel stimuli above cross-modal congruent trained stimuli (see Table 2). We also found some additional regions that showed a congruency effect in the absence of a cross-modal training effect (see Table 2). In addition, some areas showed responses that were lower for incongruent than for novel cross-modal birds. These

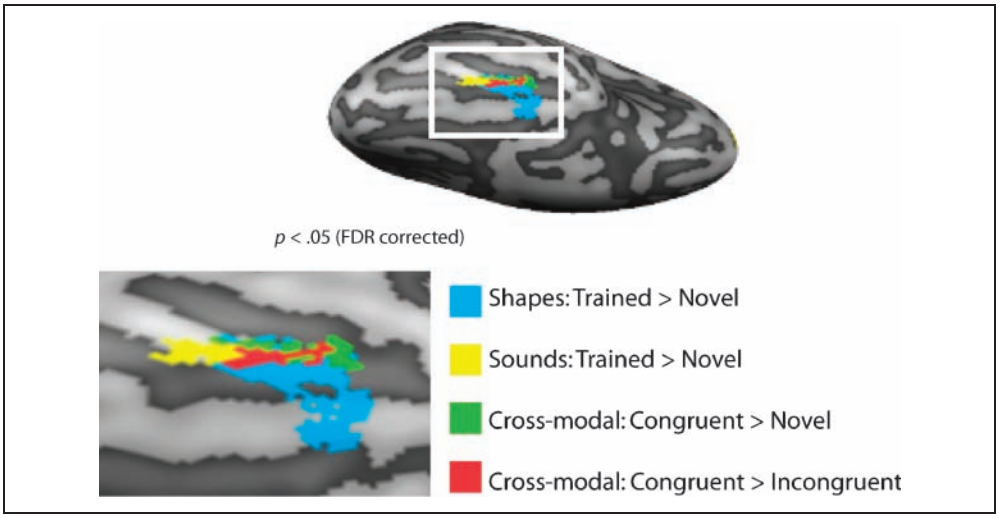

Figure 5. Right fusiform training effects. Overlap of regions in the right fusiform gyrus that show a cross-modal training effect (in green: Congruent Cross-modal Trained > Cross-modal Novel), unimodal training effects (in blue: Trained Shapes > Novel Shapes; in yellow: Trained Sounds > Novel Sounds), and a congruency effect (in red: Trained Congruent cross-modal > Trained Incongruent Cross-modal), presented at $p < .05$ corrected for display purposes.

areas were right perirhinal cortex and posterior lateral sulcus and left superior temporal gyrus, insula, supramarginal gyrus, precentral gyrus, and superior frontal sulcus.

Effects of Morph Level

To investigate the effect of morph level on the brain's responses, we collapsed over all trained conditions and tested for areas that showed greater activity for morph level 90 than for morph level 60 at $p < .01$ (FDR corrected; see Figure 7A). The areas that showed an overall effect of morph level fit nicely with those areas that responded more to trained than to novel birds and more to congruent than incongruent bird types. No areas were found that responded more to the 60% morph levels than to the 90% morph levels. For the novel bird types, we performed the same analysis, but no areas preferred the higher morph level at $p < .01$ (FDR corrected) and not even at $p < .05$ (FDR corrected) or $p < .001$ (uncorrected). We investigated responses from two areas in the right STS (Figure 7B and C), the left STS (Figure 7D), and the right occipitotemporal cortex (Figure 7E) with a MANOVA. This ROI analysis confirmed the overall effect of morph level for trained items in all these areas and not for the novel items. Interestingly, the incongruent items also showed an effect of morph level in these areas.

Effective Connectivity Analysis

We did an exploratory PPI analysis to see which areas showed greater connectivity from the left STS during presentation of cross-modal congruent birds than during presentation of cross-modal novel birds (Figure 8A). We found that the bilateral supramarginal gyrus and anterior cingulate gyrus showed more connectivity with left STS; in addition, we found a group of left-lateralized areas that included

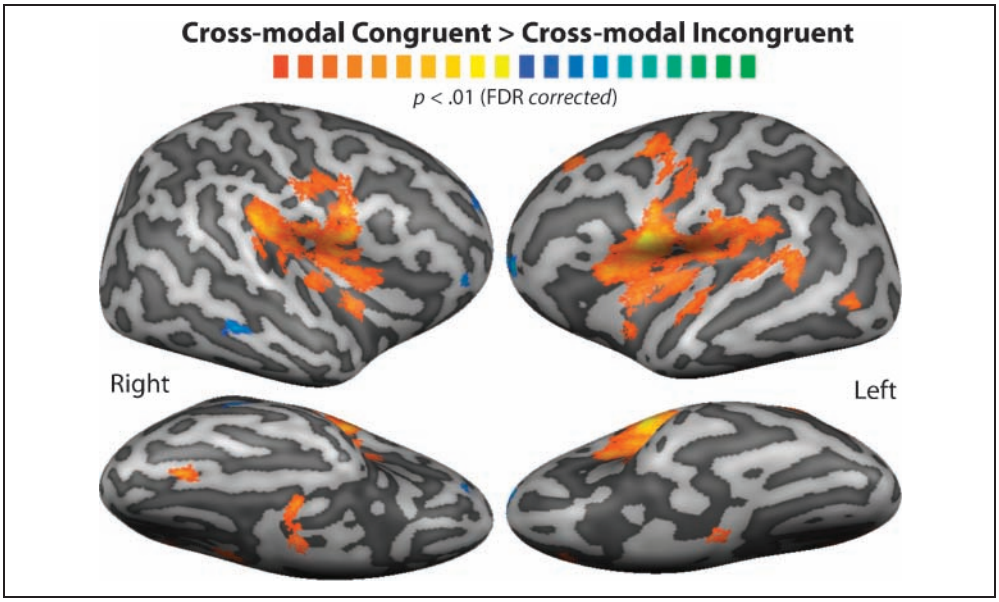

Figure 6. Congruency effects. Group-averaged activation maps from the posttraining scanning overlaid on lateral (top) and ventral (bottom) views of Talairach-normalized inflated hemispheres. In orange tones, regions that showed more activity for trained congruent cross-modal bird types compared with trained incongruent cross-modal bird types at $p < .01$ (FDR corrected). In blue, voxel populations showing more activity following presentation of trained incongruent cross-modal bird types compared with trained congruent cross-modal bird types.

inferior frontal areas, left middle frontal, and postcentral gyrus. Most interesting was that in the right occipito-temporal cortex, the right fusiform gyrus showed increased connectivity from the left STS for the cross-modal congruent bird types (Figure 8B).

DISCUSSION

In this study, we used a novel audiovisual training paradigm to investigate the formation of cross-modal object representations in the adult human brain. We trained subjects to dissociate between three highly similar cross-modal bird categories. Our behavioral results indicate that our one-back discrimination task was successful in inducing the formation of new category representations. Behavioral data from our study follow the pattern that is typical of category learning; that is, responses to stimuli that were close to the category boundary were faster and more accurately than would be expected on the basis of the physical properties of the stimuli. Even for morph ratios near the category boundary (55:45 morphs), performance exceeded 70% at the end of training. Thus, although a 55:45 exemplar of, say, bird type A had only 55% of A properties (and 45% of another bird type), it was nonetheless categorized as type A 70% of the time. This demonstrates that subjects had developed categorical perception of the bird types. Such a behavioral pattern has previously been found for training with a discrimination task (van der Linden et al., 2008; Op de Beeck et al., 2006) as well as for categorization training (van der Linden et al., 2010; Gillebert, Op de Beeck, Panis, & Wagemans, 2009; Jiang et al., 2007).

After 3 days of training, on the fourth day, the subjects were scanned. We presented them with the trained cross-modal birds in congruent and incongruent audiovisual combinations and with novel audiovisual bird categories. The subjects also listened and viewed novel and trained bird sounds and shapes in isolation. We found cross-modal training effects in frontal and temporal regions known to be involved in cross-modal object representations.

Many studies have determined that the STS plays a very important role in cross-modal integration. Anatomically the STS is conveniently located near the borders of auditory and visual association cortices. Functionally, it has been found to respond to auditory, visual, and audiovisual linguistic stimuli (van Atteveldt, Formisano, Blomert, et al., 2007; van Atteveldt, Formisano, Goebel, et al., 2007; Callan et al., 2004; van Atteveldt et al., 2004; Calvert et al., 2000) and to common shapes, sounds, and audiovisual objects (Hein et al., 2007; Taylor et al., 2006; Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, et al., 2004). We found that the STS becomes involved in cross-modal object representation after a relatively short amount of category training. What is even more important is that this training effect did not generalize to incongruent pairings of trained bird sounds and shapes. The STS did not show differential responses to incongruent bird types compared with novel bird types. This indicates that the formation of cross-modal representations was meaningful, namely, restricted to those combinations of sounds and shapes that were associated together during category training and did not just occur for any combination of familiar trained sounds and shapes. Congruency effects have been found before in the left STS (Calvert et al., 2000). However, the reversed effect has also been found in the STS during active matching (Hocking & Price, 2008; Taylor et al., 2006).

Although there seems great consensus that the STS is a site for cross-modal integration, it is also possible that the STS is involved in integrating or associating information regardless of modality. Recently, it was found that the STS responded in equal amounts to visual–visual, auditory–auditory, and audiovisual matching (Hocking & Price, 2008). In addition, in a study where subjects learned associations between cross-modal stimuli that were presented segregated in time, the STS increased its responsiveness as learning progressed for visual–visual and audiovisual associations (Tanabe, Honda, & Sadato, 2005). In a previous study, we also found that the STS is involved in learned associations between birds from different perceptual categories (van der Linden et al., 2010). The results from the present study further support the theory that the STS is involved in associative learning or linking different types of information regardless of modality. In general, one can say that repeated simultaneous presentation of sound and image during training results in the association of these unimodal representations. It is likely that our training paradigm with morphed cross-modal birds made the association of sound and shape extra salient. Especially for the more difficult birds around the category border, combining the information of both modalities probably provided stronger clues to category membership than each modality in isolation would have provided. Therefore, training strengthened the association between sound and shape representation, and the successful association of these unimodal representations into a congruent cross-modal category can explain the cross-modal training and congruency effects in the STS.

Another region that showed a cross-modal training effect was the right fusiform gyrus. We found that cross-modal training with the birds resulted in increased activity for cross-modal birds with congruent sounds and shapes as compared with cross-modal novel birds. Importantly, this training-related increase in responses was not present for incongruent trained bird types. Moreover, the response to cross-modal congruent bird types was larger than the response to incongruent trained bird types. This fits the results of Naumer et al. (2009) who also report a congruency effect for trained cross-modal nonsense objects. Interestingly, in our study, in the right fusiform gyrus a training-related increase was present for trained shapes in the absence of sounds as well as for trained sounds in the absence of shapes. The finding of a cross-modal training effect combined with a training effect for bird shapes

Table 2. Regions Showing a Congruency Effect

Area	*x*	*y*	*z*	*mm*^{3}	*CCT > CN*	*CIT > CN*	*CCT > CIT*
Congruent > Incongruent							
L perirhinal cortex	−9	−32	−17	511	1.18	−2.04	3.87*
R perirhinal cortex	17	−21	−19	678	1.19	−3.12**	4.48***
R posterior fusiform G	31	−55	−16	365	1.28	−1.55	3.27**
L lateral occipital G	−41	−63	3	218	1.06	−1.75	2.53****
L cuneus	−7	−74	13	215	3.01**	−0.33	3.96*
R posterior cingulate G	7	−38	7	392	1.10	−1.84	2.53****
L posterior cingulate G	−8	−60	10	557	2.93****	−0.02	4.10***
L parieto-occipital S	−17	−67	16	155	2.32****	0.08	3.36*
R parieto-occipital S	10	−58	19	240	2.01	−0.62	2.65****
L anterior cingulate	−2	22	6	504	2.34****	−0.36	2.89****
R anterior cingulate G	9	−10	39	959	1.50	−1.05	3.02**
L insula	−34	−16	11	2501	2.36****	−2.43****	4.14***
	−36	−6	−8	220	0.78	−1.86	2.67****
R insula	35	−12	0	412	0.95	−1.67	3.63*
	34	−7	16	1228	2.31****	−1.18	3.32*
L superior temporal S	−46	−48	9	324	2.69****	−0.56	3.14**
L superior temporal G	−56	−49	14	1087	2.76****	−0.79	2.91****
	−57	−37	14	1027	1.31	−2.02	2.97**
	−54	−17	10	860	1.23	−2.28****	2.91****
R superior temporal G	37	−21	9	475	2.15****	−1.70	3.39*
L supramarginal G	−49	−44	27	1168	2.13	−1.29	3.04**
	−55	−29	22	2722	1.95	−1.79	4.05*
	−38	−28	21	2452	1.83	−2.42****	4.41***
R supramarginal G	52	−34	19	346	1.18	−2.09	3.56*
	52	−34	33	1036	2.48****	−0.58	2.70****
L inferior frontal G	−34	4	17	1615	2.21****	−0.54	3.27**
L precentral G	−52	−5	14	3314	1.77	−2.71****	4.10***
	−43	−13	48	1458	1.88	−1.27	4.54***
R precentral G	52	−8	30	2611	3.09**	−1.19	4.02*
	51	−8	15	2812	2.47****	−1.86	3.97*
L postcentral G	−44	−15	33	1259	1.81	−2.09	3.70*
R posterior lateral S	43	−29	22	3032	1.72	−2.24****	3.44*
L superior frontal G	−9	−15	46	4633	1.79	−1.70	4.01*
L superior frontal S	−22	19	45	894	2.53****	−2.24****	2.54****

Table 2. (*continued*)

Area	*x*	*y*	*z*	mm^3	*CCT > CN*	*CIT > CN*	*CCT > CIT*
Incongruent > Congruent							
R middle temporal G	57	−38	−4	261	−2.04	0.81	−2.40****
R inferior frontal G	44	38	−1	194	−2.51****	0.59	−2.23****
L middle frontal G	−26	53	11	827	−1.51	1.78	−2.59****

Mean Talairach coordinates, volume in cubic millimeters, and averaged *t* values for regions showing a cross-modal congruency effect at $p < .01$, FDR corrected. *t* values for the paired *t* tests ($df = 15$) on the subject-averaged beta weights of the congruency effect are presented: Cross-modal Congruent Trained > Cross-modal Incongruent Trained (CCT > CIT). In addition, we present *t* values obtained for the cross-modal training effect: Cross-modal Congruent Trained > Cross-modal Novel (CCT > CN).

L = left; R = right; G = gyrus; S = Sulcus.

*$p < .005$.

**$p < .01$.

***$p < .001$.

****$p < .05$.

presented in isolation fits well with a previous fMRI study in which we found increased fusiform responses for bird types that subjects successfully learned to visually dissociate (van der Linden et al., 2008). Increased activity in the fusiform gyrus has also been found after subjects became proficient in individuating a homogeneous set of nonsense objects (Gauthier et al., 1999). In addition, larger fusiform responses were observed in individuals that were highly skilled in recognizing a particular class of objects such as birds, cars, or Lepidoptera (butterflies and moths) (Xu, 2005; Rhodes et al., 2004; Gauthier et al., 2000). The fact that the right middle fusiform gyrus showed no training effect for incongruent cross-modal bird stimuli also fit with our previous finding that the right fusiform gyrus showed only increased responsiveness for birds for which a meaningful representation had been formed and not for birds to which the subjects were exposed in an equal amount but for which they were hindered in forming a representation of the categories (van der Linden et al., 2008). It is likely that the fusiform gyrus is involved in coding for the visual features of the bird types that were informative during cross-modal training.

Because the fusiform is part of the ventral visual stream, finding unimodal auditory training effects in the fusiform gyrus is somewhat surprising. However, Beauchamp, Lee, et al. (2004) also reported auditory activation in the ventral visual stream for sounds of common objects presented in isolation. Responses in the fusiform gyrus seem to emerge when sounds are presented for which a visual association exists. During voice recognition, the fusiform gyrus showed larger responses for voices associated with a familiar face than for unfamiliar voices (von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). The same was found for voices that were associated with a face as a result of training (von Kriegstein & Giraud, 2006). In our study, hearing the sound of a bird that was trained might have activated the associated visual representation. Such a representation did not exist for novel birds; therefore, novel bird sounds did not activate the fusiform gyrus. This gave rise to the observed auditory training effect in the right fusiform gyrus. Activation of the visual representation of a bird by its sound could also explain why the fusiform gyrus shows a congruency effect. In line with this reasoning is the finding of tighter connection strength for cross-modal trained birds than for cross-modal novel birds between the left STS and the right fusiform gyrus. This could reflect that training increased top–down influence of the STS on the right fusiform gyrus. Therefore, when presented with a congruent cross-modal bird, both its sound, via feedback connections of the superior temporal sulcus into the fusiform gyrus, and shape activated the newly formed visual representation of the bird. This might boost activation in this area. For incongruent birds, the shape might have activated the visual representation in the fusiform, but the combined sound did not match this representation; therefore, no increase in activation was observed.

The inferior and the middle frontal gyri showed the reverse effect of the temporal areas and responded more to incongruent cross-modal birds compared with congruent cross-modal bird types. This result corroborates with other studies (Hein et al., 2007; Belardinelli et al., 2004). Rather than being involved in cross-modal binding, the inferior frontal cortex is linked to semantic retrieval (Martin & Chao, 2001; Wagner, Pare-Blagoev, Clark, & Poldrack, 2001). Presenting subjects with incongruent cross-modal stimuli could have reflected increased load on semantic memory because retrieval of a semantic representation was unsuccessful. This failure to retrieve a semantic representation could also explain why we found larger responses to novel birds compared with trained congruent birds in this area. Our findings of temporal areas showing congruency effects and of frontal areas showing the reversed effect are the same pattern that was recently described in a review article (Doehrmann & Naumer, 2008) that evaluated the role of semantics on audiovisual integration in frontal and temporal regions.

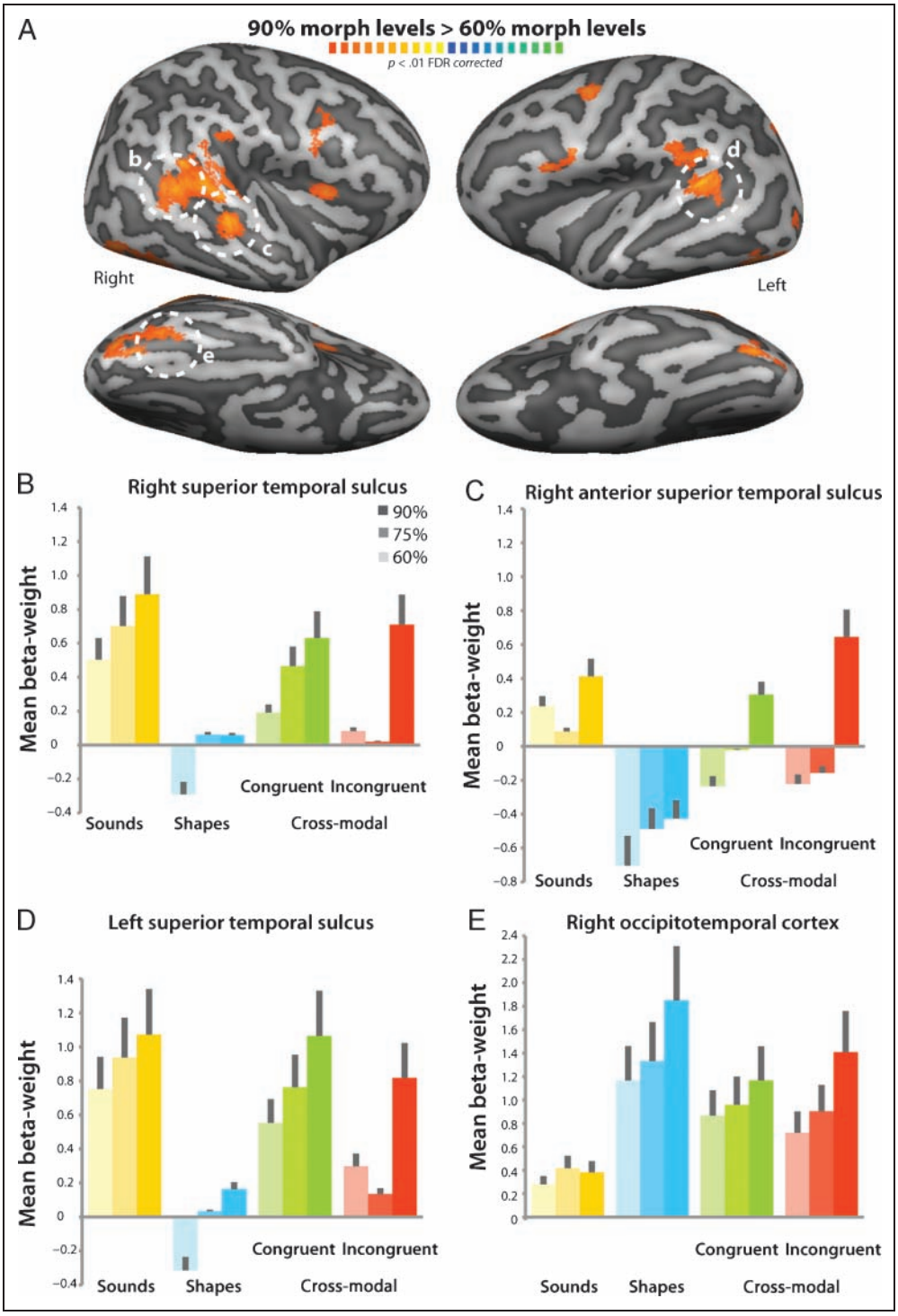

Figure 7. Effects of morph level. (A) Shown in orange colors are areas that responded more to 90% morph levels than to 60% morph levels of the trained bird types at $p < .01$ (FDR corrected). The activations are overlaid on lateral (top) and ventral (bottom) views of Talairach-normalized inflated hemispheres. Plots show the voxel-averaged mean beta-weights in (B) the right STS, (C) the right anterior STS, (D) the left STS, and (E) the right occipito-temporal cortex. Shown are the averaged responses for the trained bird types for the unimodal bird stimuli (sounds in yellow and shapes in blue) and cross-modal bird stimuli (green for congruent and red for incongruent stimuli). Color saturation represents the morph levels, the most saturated color represents the 90% morph level, and the least saturated color the 60% morph level. Error bars represent *SEM*.

We found that several areas, among which the STS and the occipito-temporal cortex, showed an effect of morph level. Responses were greater to birds with a higher percentage morph level. These areas were for the most part the same areas that preferred trained over novel and congruent over incongruent bird types. The effect of morph level is experience dependent; we found it only for the trained bird types and not for novel birds. Interestingly, the incongruent cross-modal birds also showed an effect of morph level. This indicates that although the incongruent recombinations of trained sounds and birds activated some general representation of the birds and that this representation was influenced by categorization training, the response is higher to those birds that are further away from the category boundary.

Our analysis of effective connectivity showed that several areas showed increased connectivity with the left STS as a result of training. These areas included the left frontal

areas, the right occipito-temporal cortex, and the bilateral supramarginal gyrus. These areas are overlapping those areas that showed cross-modal congruency and training effects. We already discussed the putative roles of frontal and occipito-temporal areas in cross-modal processing. The supramarginal gyrus has been recently found to be involved in successful category learning of sounds (Liebenthal et al., 2010; Desai, Liebenthal, Waldron, & Binder, 2008). In addition, faster learners of nonnative speech sounds have greater white matter volume in bilateral supramarginal gyrus than slow learners (Golestani, Paus, & Zatorre, 2002). Taken together with the findings from the present study, this suggests that the supramarginal gyrus is involved in learning auditory categories.

One particular concern in this study is the role of attention on the processing of the stimuli. Like in many other studies (Naumer et al., 2009; Hein et al., 2007; Belardinelli et al., 2004; van Atteveldt et al., 2004; Calvert et al., 2000), we used a passive task and blocked presentation. An alternative explanation, therefore, might be that the congruency effect is attributed to differences in attention. However, when van Atteveldt, Formisano, Goebel, et al. (2007) compared passive blocked presentations of cross-modal stimuli with a passive event-related paradigm, the congruency effects did not disappear. When comparing the passive paradigm with an active matching paradigm, they found that the congruency effects disappeared during active matching and even resulted in incongruency effects in several other brain regions. In addition, novel stimuli in all modalities are usually associated with higher attentional engagement and thus higher BOLD responses (Downar, Crawley, Mikulis, & Davis, 2002). Therefore, one can expect that novel birds and new recombinations of trained sounds and shapes, that is, the incongruent bird types, would show larger responses than the trained birds. However, in our study, there were very few regions that preferred novel or incongruent stimuli.

To summarize, with this caveat in mind, the present study revealed plasticity in the adult human brain resulting from the successful association of bird sounds and bird shapes into coherent cross-modal categories. Cross-modal training and congruency effects revealed the representation of these meaningful cross-modal categories. These cross-modal training effects indicate that the cortical representation of audiovisual object categories is experience dependent, being more involved in processing trained bird types than similar novel birds. Moreover, this representation is category specific; it is based on learned associations between sounds and shapes that define a category. Learning did not generalize to incongruent combinations of trained sounds and shapes. We observed cross-modal, auditory, and visual training effects in the right fusiform gyrus that did not generalize to incongruent combinations of sound and shape. Given the involvement of the right fusiform gyrus in learning to categorize visual objects (van der Linden et al., 2008; Gauthier et al., 1999), we conclude that the right fusiform gyrus was involved in the visual representation of the learned bird shapes. Another region showing cross-modal training and congruency effects was the left STS. Rather than being just a binding site for visual and auditory properties of objects, the STS is involved in the representation of associated objects (van der Linden et al., 2010; Hocking & Price, 2008; Tanabe et al., 2005). We conclude that this area was involved in the formation of new meaningful links between sound and shapes of birds. The present study thus provides the first evidence that the adult human brain is indeed plastic enough to learn new cross-modal categories by the associations of sounds and shapes. Moreover, the combination of sound and shapes that define a

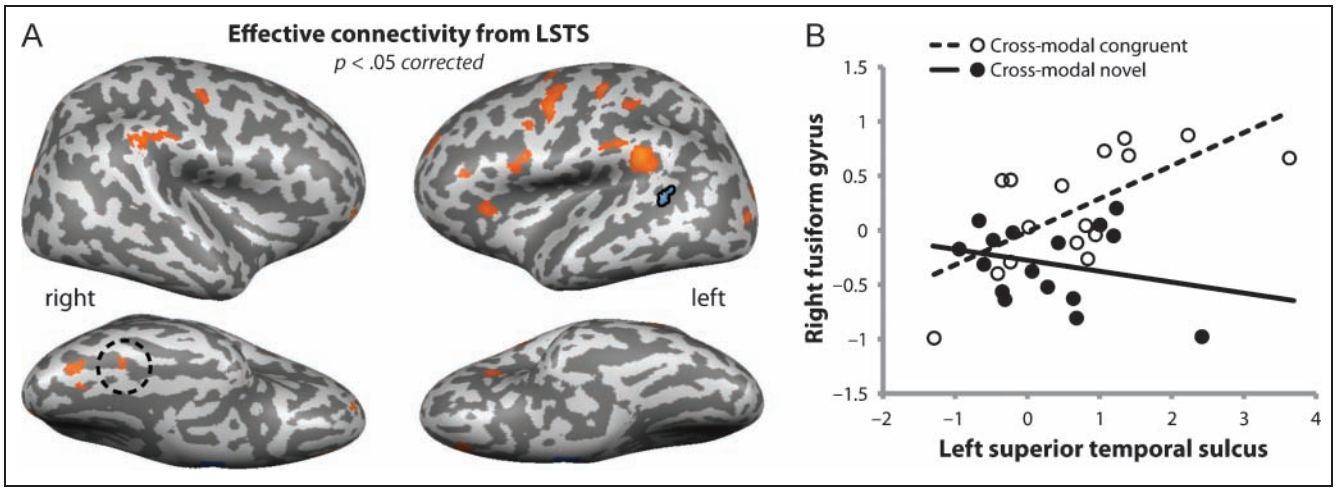

Figure 8. Effective connectivity analysis. (A) Areas that show increased connectivity at $p < .05$ (FDR corrected) from the seed region in the left STS (represented in blue with a black outline) for cross-modal congruent birds as compared with cross-modal novel birds are presented in orange colors. (B) Scatter plots of the correlation of activity (mean beta-weights) between the right fusiform gyrus on the *y*-axis and the left STS on the *x*-axis. Black dots and the solid black line represent the cross-modal novel birds ($R^2 = .07$) and the open dots with the dotted line represent the cross-modal congruent birds ($R^2 = .45$).

category is crucial for the formation of cortical cross-modal representations.

Acknowledgments

This research was supported by a grant from the Netherlands Organisation for Scientific Research (NWO 400-03-338).

Reprint requests should be sent to Marieke van der Linden, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands, or via e-mail: mail@mariekevanderlinden.com.

REFERENCES

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human cross-modal identification and object recognition. *Experimental Brain Research, 166,* 559–571.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7,* 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Belardinelli, M. O., Sestieri, C., Matteo, R., Delogu, F., Gratta, C., Ferretti, A., et al. (2004). Audio-visual crossmodal interactions in environmental perception: An fMRI investigation. *Cognitive Processing, 5,* 167–174.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57,* 289–300.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience, 16,* 4207–4221.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16,* 805–816.

Calvert, G., & Lewis, J. W. (2004). Hemodynamic studies of audiovisual interactions. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 483–502). Cambridge, MA: MIT Press.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7,* 460–467.

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience, 20,* 1174–1188.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research, 1242,* 136–150.

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2002). A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities. *Journal of Neurophysiology, 87,* 615–620.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291,* 312–316.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience, 23,* 5235–5246.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage, 6,* 218–229.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3,* 191–197.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience, 2,* 568–573.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage, 15,* 870–878.

Gillebert, C. R., Op de Beeck, H. P., Panis, S., & Wagemans, J. (2009). Subordinate categorization enhances the neural selectivity in human object-selective cortex for fine shape differences. *Journal of Cognitive Neuroscience, 21,* 1054–1064.

Golestani, N., Paus, T., & Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron, 35,* 997–1010.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10,* 14–23.

Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience, 27,* 7881–7887.

Hocking, J., & Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex, 18,* 2439–2449.

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., VanMeter, J., & Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron, 53,* 891–903.

Kriegeskorte, N., & Goebel, R. (2001). An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes. *Neuroimage, 14,* 329–346.

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., & Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex*, *20,* 2958–2970.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11,* 194–201.

Moore, C. D., Cohen, M. X., & Ranganath, C. (2006). Neural mechanisms of expert skills in visual working memory. *Journal of Neuroscience, 26,* 11187–11196.

Naumer, M. J., Doehrmann, O., Muller, N. G., Muckli, L., Kaiser, J., & Hein, G. (2009). Cortical plasticity of audio-visual object representations. *Cerebral Cortex, 19,* 1641–1653.

Olivetti Belardinelli, M., Sestieri, C., Matteo, R., Delogu, F., Gratta, C., Ferretti, A., et al. (2004). Audio-visual crossmodal interactions in environmental perception: An fMRI investigation. *Cognitive Processing, 5,* 167–174.

Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., & Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *Journal of Neuroscience, 26,* 13025–13036.

Rhodes, G., Byatt, G., Michie, P. T., & Puce, A. (2004). Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience, 16,* 189–203.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-Dimensional proportional system: An approach to medical cerebral imaging*. New York: Thieme Medical Publishers.

Tanabe, H. C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *Journal of Neuroscience, 25,* 6409–6418.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 8239–8244.

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

van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex, 17,* 962–974.

van Atteveldt, N. M., Formisano, E., Goebel, R., & Blomert, L. (2007). Top–down task effects overrule automatic multisensory responses to letter-sound pairs in auditory association cortex. *Neuroimage, 36,* 1345–1360.

van der Linden, M., Murre, J. M. J., & van Turennout, M. (2008). Birds of a feather flock together: Experience-driven formation of visual object categories in the human brain. *PLoS ONE, 3,* e3995.

van der Linden, M., van Turennout, M., & Indefrey, P. (2010). Formation of category representations in superior temporal sulcus. *Journal of Cognitive Neuroscience, 22,* 1270–1282.

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4,* e326.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17,* 367–376.

Wagner, A. D., Pare-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron, 31,* 329–338.

Weisberg, J., van Turennout, M., & Martin, A. (2007). A neural system for learning about object function. *Cerebral Cortex, 17,* 513–521.

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cerebral Cortex, 15,* 1234–1242.