Behavioral/Systems/Cognitive

Direct Structural Connections between Voice- and Face-Recognition Areas

Helen Blank, Alfred Anwander, and Katharina von Kriegstein
Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

Currently, there are two opposing models for how voice and face information is integrated in the human brain to recognize person identity. The conventional model assumes that voice and face information is only combined at a supramodal stage (Bruce and Young, 1986; Burton et al., 1990; Ellis et al., 1997). An alternative model posits that areas encoding voice and face information also interact directly and that this direct interaction is behaviorally relevant for optimizing person recognition (von Kriegstein et al., 2005; von Kriegstein and Giraud, 2006). To disambiguate between the two different models, we tested for evidence of direct structural connections between voice- and face-processing cortical areas by combining functional and diffusion magnetic resonance imaging. We localized, at the individual subject level, three voice-sensitive areas in anterior, middle, and posterior superior temporal sulcus (STS) and face-sensitive areas in the fusiform gyrus [fusiform face area (FFA)]. Using probabilistic tractography, we show evidence that the FFA is structurally connected with voice-sensitive areas in STS. In particular, our results suggest that the FFA is more strongly connected to middle and anterior than to posterior areas of the voice-sensitive STS. This specific structural connectivity pattern indicates that direct links between face- and voice-recognition areas could be used to optimize human person recognition.

Introduction

Successful face-to-face communication relies on decoding sensory information from multiple modalities, such as the visual face and the auditory voice. How does our brain integrate this multisensory information to recognize a person? Models for person recognition make two opposing predictions: The conventional model (Fig. 1*A*) assumes that faces and voices are processed separately until the person is identified at a supramodal level of person recognition (Bruce and Young, 1986; Burton et al., 1990; Ellis et al., 1997). An alternative model (Fig. 1*B*) extends this view and posits that information can also be combined by direct interactions between voice- and face-processing areas. Such a direct integration of information would provide useful constraints to resolve ambiguity in noisy input (von Kriegstein and Giraud, 2006; von Kriegstein et al., 2008). This would be advantageous for optimizing person recognition under natural conditions, e.g., in noisy environments or under less than optimal viewing or hearing conditions.

Voice-sensitive areas have been localized along the superior temporal sulcus (STS) (Belin et al., 2000; von Kriegstein and Giraud, 2004). Posterior areas of the STS are more involved in acoustic processing and more anterior areas are responsive to voice identity (Belin and Zatorre, 2003; von Kriegstein et al., 2003; von Kriegstein and Giraud, 2004; Andics et al., 2010). Face-sensitive areas are located in occipital gyrus, fusiform gyrus, and anterior inferior temporal lobe (Kanwisher et al., 1997; Kriegeskorte et al., 2007; Rajimehr et al., 2009). The area that is most selective and reliably activated for faces is the fusiform face area (FFA) (Kanwisher et al., 1997). It is not only involved in the processing of facial features, but also in face-identity recognition (Sergent et al., 1992; Eger et al., 2004; Rotshtein et al., 2005).

Several recent behavioral, electrophysiological, and functional magnetic resonance imaging (fMRI) studies support a model with direct interactions between voice-sensitive STS and the FFA (Fig. 1*B*) (Sheffert and Olson, 2004; von Kriegstein et al., 2005, 2008; von Kriegstein and Giraud, 2006; Föcker et al., 2011). A prerequisite for such a model are direct structural connections between these auditory and visual areas. It is currently unknown whether such structural connections exist. Here we combine fMRI and diffusion magnetic resonance imaging (dMRI) to test this.

Direct structural connections between voice- and face-sensitive areas would be in line with recent developments in multisensory research suggesting that information from different modalities interacts already at relatively early processing stages (Ghazanfar and Schroeder, 2006; Kayser and Logothetis, 2007; Driver and Noesselt, 2008; Cappe et al., 2010; Kayser et al., 2010; Klinge et al., 2010). Early integration based on direct connections would provide useful constraints for possible interpretations of ambiguous sensory input (von Kriegstein et al., 2008). This account would also integrate well with theories of multisensory processing (Ernst and Banks, 2002) and with general theories of brain function (Friston, 2005, 2010; Kiebel et al., 2008).

Received April 27, 2011; revised June 22, 2011; accepted July 14, 2011.

Author contributions: H.B. and K.v.K. designed research; H.B. and K.v.K. performed research; A.A. contributed unpublished reagents/analytic tools; H.B. and A.A. analyzed data; H.B. and K.v.K. wrote the paper.

This work was funded by a Max Planck Research Group grant to K.v.K. We thank Bernhard Comrie for providing the high-quality stimulus recording environment; Nathalie Fecher, Sven Grawunder, and Peter Froehlich for their help with recordings of the stimulus material; and Stefan Kiebel for helpful discussions.

The authors declare no conflict of interest.

Correspondence should be addressed to Helen Blank, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1A, 04103 Leipzig, Germany. E-mail: hblank@cbs.mpg.de.

DOI:10.1523/JNEUROSCI.2091-11.2011

 0270-6474/11/3112906-10$15.00/0

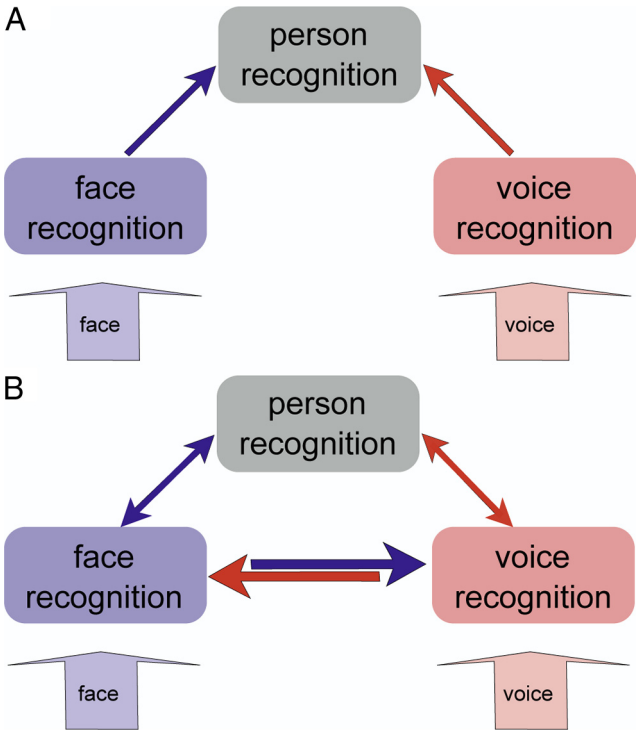

Figure 1. Two models for person recognition. ***A***, Unisensory information is integrated at a supramodal stage of person recognition (Burton et al., 1990; Ellis et al., 1997). ***B***, Unisensory information can also be integrated using direct reciprocal interactions between sensory areas (von Kriegstein et al., 2005; von Kriegstein and Giraud, 2006). Arrows indicate possible structural connections between areas.

Materials and Methods

Subjects

Twenty-one healthy volunteers (mean age, 26.9 years; age range, 23–34 years; all right-handed [assessed with the Edinburgh questionnaire (Oldfield, 1971)]; 10 female) participated in our study. Written informed consent was collected from all participants according to procedures approved by the Research Ethics Committee of the University of Leipzig. Two subjects were excluded from the analysis: the first because of difficulties with acquiring the field-map during fMRI and the second because he did not follow the task instructions. Furthermore, one subject's behavioral results for the second functional localizer (see Localizer 2: person and object recognition, below) had to be excluded due to intermittent technical problems with the response box.

Stimuli

Stimuli consisted of videos (with and without audio-stream) and auditory-only files. Stimuli were created by recording three male speakers (22, 23, and 25 years old) and three mobile phones. All recordings were done in a soundproof room under constant luminance conditions. Videos were taken of the speakers' faces and of a hand operating the mobile phones. Speech samples of each speaker included semantically neutral, phonologically, and syntactically homogeneous five-word sentences (e.g., "Der Junge trägt einen Koffer."/"The boy carries a suitcase."), two-word sentences (the pronoun "er"/"he" and a verb; e.g., "Er kaut."/"He chews."), and single words (e.g., "Dichter"/"poet"). Key tone samples of each mobile phone included several different sequences of two to nine key presses per sequence. Videos were recorded with a digital video camera (Legria HF S10 HD-Camcorder; Canon). High-quality auditory stimuli were simultaneously recorded with a condenser microphone [TLM 50 (Neumann); preamplifier, Mic-Amp F-35 (Lake People); soundcard, Power Mac G5 (Apple); 44.1 kHz sampling rate and 16 bit resolution] and the software Sound Studio 3 (Felt Tip).

The auditory stimuli were postprocessed using Matlab (version 7.7; MathWorks) to adjust overall sound level. The audio files of all speakers and mobile phones were adjusted to a root mean square of 0.083. All videos were processed and cut in Final Cut Pro (version 6, HD; Apple), converted to mpeg format, and presented at a size of 727 × 545 pixels.

Procedure

All subjects participated in two fMRI localizer scans, a dMRI scan, and a structural T1 scan. The fMRI localizer scans were performed on a different day than the dMRI and T1 scans (Fig. 2*A*).

Functional localizers

The location of the FFA and the voice-sensitive regions in STS differs considerably between subjects (Kanwisher et al., 1997; Belin et al., 2002). We therefore localized the areas of interest on the single-subject level. We used a standard fMRI contrast to localize the voice-sensitive areas in STS (von Kriegstein et al., 2003; von Kriegstein and Giraud, 2004). To locate the FFA, we used two contrasts. First, we used the standard contrast to localize the FFA [viewing faces vs viewing objects (Kanwisher et al., 1997)]. Second, because we were specifically interested in localizing the area processing identity, we used a more specific contrast that shows FFA responses during auditory-only voice recognition after face-identity learning (von Kriegstein et al., 2005, 2006, 2008; von Kriegstein and Giraud, 2006).

Training

Before fMRI scanning, all participants were trained to identify the speakers and the mobile phones. The training served to induce FFA responses during auditory-only voice recognition in localizer 1 (see Localizer 1: voice and speech recognition, below) and to train the subjects so that they could perform the recognition tasks in localizer 2 (see Localizer 2: person and object recognition, below). The three speakers were learned by watching videos of their faces and hearing their voices saying 36 five-word sentences. Subjects also learned to recognize the three mobile phones by audiovisual videos showing a hand pressing keys. Thirty-six sequences with different numbers of key presses were used. After learning, participants were tested on their recognition performance. In this test, subjects first saw silent videos of a person (or mobile phone) and subsequently listened to a voice (or key tones). They were asked to indicate whether the auditory voice (or key tone) belonged to the face (or mobile phone) in the video. Subjects received feedback about correct, incorrect, and too slow responses. The training, including learning and test, took 25 min. Training was repeated twice for all participants. If a participant performed <80% correct after the second repetition, the training was repeated a third time.

Localizer 1: voice and speech recognition

This experiment was used to localize the voice-sensitive areas in the STS and the FFA in response to auditory-only voice recognition (von Kriegstein and Giraud, 2006; von Kriegstein et al., 2008). The experiment contained a voice recognition and a speech recognition control condition using the same stimuli (Fig. 2*B*). Auditory-only two-word sentences were presented in blocks of 21.6 s duration. Each block was followed by a silent period in which a fixation cross was shown for 13 s. Before each block, participants received the written on-screen instruction to perform either the voice- or speech-recognition task. At the same time, they were presented with an auditory target sentence spoken by one of the three previously learned speakers. In the voice task, subjects were asked to decide whether a sentence was spoken by the target speaker or not, independent of which specific sentence was said. In the speech task, subjects were asked to decide for each sentence whether it was the target sentence or not, independent of which speaker it said. Responses were made via button press. Note that during both conditions, we presented exactly the same set of auditory-only stimuli. Stimuli were presented and responses recorded using Presentation software 14.1 (http://nbs.neuro-bs.com). Conditions were split into 18 blocks (nine blocks of voice recognition and nine blocks of speech recognition) presented in random order within and across conditions. Each block contained 12 items (three two-word sentences were repeated four times). Four items in each block were targets. The stimuli within a block were chosen to sound similar [e.g., "Er kauft.", "Er kaut.", "Er klaut." (in English: "He shops.", "He chews.", "He

steals.")] to approximately match the difficulty of the speech to the voice task. Total scanning time for this localizer was 11.46 min.

Localizer 2: person and object recognition
This fMRI design (Fig. 2*C*) was used to localize the FFA with the standard contrast visual "face stimuli versus object stimuli" (Kanwisher et al., 1997). By combining an object (a mobile phone) and human hand in the control stimulus, we used a combination of stimuli frequently used in FFA localizers (Kanwisher et al., 1997; Fox et al., 2009; Berman et al., 2010). These two conditions were embedded in a more complex experiment that was designed to address a different research question. The results will be reported elsewhere. For completeness, we describe the full experimental setup. The experiment was a 2 × 2 × 2 factorial design [2 categories (persons and mobile phones) × 2 modalities (auditory stimulus first vs visual stimulus first within a single trial) × 2 tasks (recognition and matching task)]. In the person condition, each trial consisted of an auditory-only presentation of the voice and a separate visual-only video of a talking face. The stimuli were taken from the audiovisually recorded single words. In the mobile phone condition, each trial consisted of an auditory-only presentation of the key tone and a separate visual-only video of the mobile phone. In the identity recognition task, participants were requested to indicate via key press whether the visual face (mobile phone) and the auditory-only voice (key tone) belonged to the same person (mobile phone) or not. In the matching task, exactly the same stimulus material was shown. Participants were requested to indicate via button press whether the visual-only presented word (number of key tones) matched the auditory-only presented word (number of key tones) or not. Matching here means that the auditory-only presented word (number of key tones) was exactly the same word (number of key tones) as in the visual-only presentation. In each trial, the first stimulus was presented for ~1.6 s, depending on the duration of the stimuli. The second stimulus was surrounded by a blue frame indicating the response phase (2.3 s). If the first stimulus was auditory-only, the second stimulus was visual-only and vice versa. Stimuli were separated by a fixation cross that was presented for a jittered duration with an average of 1.6 s and a range of 1.2–2.2 s. One third of the events were null events of 1.6 s duration randomly presented within the experiment. The trials for each condition were grouped into sections of 12 trials (84 s) to minimize time spent on instructing the subjects which task to perform. Sections were presented in semirandom order, not allowing neighboring sections of the same condition. All sections were preceded by a short instruction about the task. The written words "person" and "mobile phone" indicated the identity-recognition task, whereas "word" and "key press" indicated the matching task. Each instruction was presented for 2.7 s. The whole experiment consisted of two 16.8 min scanning sessions. Before the experiment, subjects received a short familiarization with all tasks.

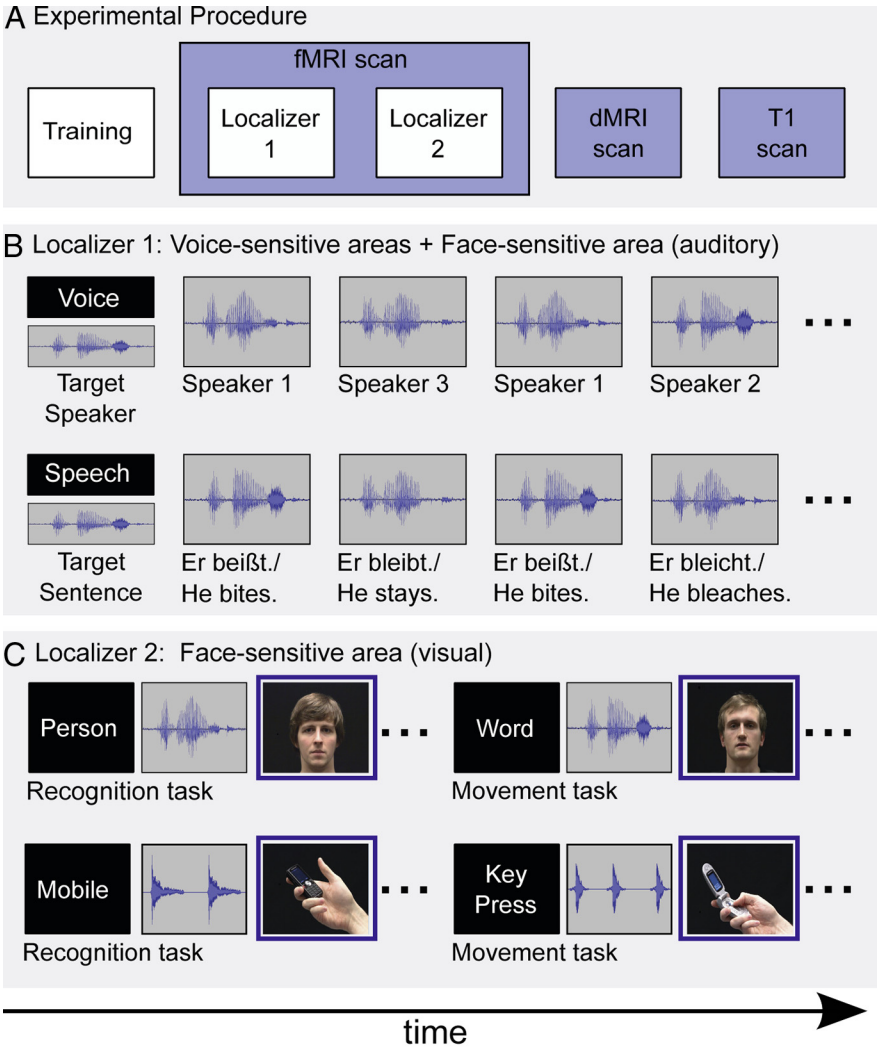

Figure 2. Experimental procedure and design of the two functional localizers. ***A***, Experimental procedure. All subjects participated in a training session before the two functional MRI localizer scans. In addition, we acquired, on a different day, a dMRI and a structural T1 scan. ***B***, Localizer 1: block design to localize voice-sensitive areas in STS and responses to voices in the FFA. At the beginning of a block, subjects were instructed to perform a speech- or voice-recognition task on auditory sentences. Subjects decided for each sentence whether it was spoken by the target speaker (voice task) or whether the content of the sentence matched the target sentence (speech task). ***C***, Localizer 2: event-related design to localize face-sensitive areas in fusiform gyrus (FFA). To localize the FFA, we used a contrast of visual faces > visual mobile phones. For details, see Material and Methods.

Image acquisition

Three different kinds of images were acquired: functional images, diffusion-weighted images, and structural T1-weighted images. All images were acquired on a 3T Siemens Tim Trio MR scanner (Siemens Healthcare).

Functional MRI sequence
For the functional localizers, a gradient-echo EPI (echo planar imaging) sequence was used (TR, 2.79 s; TE, 30 ms; flip angle, 90°; 42 slices, whole-brain coverage; acquisition bandwidth, 116 kHz; 2 mm slice thickness; 1 mm interslice gap; in-plane resolution, 3 × 3 mm). Geometric distortions were characterized by a B0 field-map scan. The field-map scan consisted of gradient-echo readout (24 echoes; interecho time, 0.95 ms) with standard 2D phase encoding. The B0 field was obtained by a linear fit to the unwrapped phases of all odd echoes.

Structural MRI sequences
Diffusion MRI. For the analysis of the anatomical connectivity, we used a dMRI sequence providing high angular resolution diffusion imaging data. These data were acquired using a 32-channel coil with a twice-refocused spin echo EPI sequence (TE, 100 ms; TR, 12 s; image matrix, 128 × 128; FOV, 220 × 220 mm) (Reese et al., 2003), providing 60 diffusion-encoding gradient directions with a b-value of 1000 s/mm^2.

Table 1. Coordinates of voice- and face-sensitive regions in MNI space and associated maximum *Z*-statistic values

	Crossmodal FFA		Visual FFA		Posterior STS		Middle STS		Anterior STS	
Subjects	Coordinates	*Z*	Coordinates	*Z*	Coordinates	*Z*	Coordinates	*Z*	Coordinates	*Z*
1	—	—	39, −43, −17	3.37	72, −37, 13	1.79	—	—	—	—
2	30, −58, −20	2.11	42, −37, −17	2.38	69, −37, 10	3.42	60, −10, −8	2.04	54, 8, −5	2.88
3	36, −49, −17	2.41	—	—	48, −53, 13	3.40	69, −10, −17	2.60	57, 11, −26	2.49
4	33, −64, −5	2.60	48, −55, −23	2.03	60, −37, 7	3.14	66, −7, −11	1.89	60, 8, −14	1.96
5	27, −43, −11	2.61	—	—	69, −34, 7	1.88	57, −25, −23	2.43	—	—
6	—	—	39, −55, −23	4.52	—	—	72, −19, −17	2.08	—	—
7	—	—	42, −58, −20	3.44	—	—	—	—	54, 5, −2	2.52
8	42, −52, −20	3.06	—	—	60, −37, 16	3.25	66, −16, −17	2.34	51, 17, −14	1.96
9	36, −49, −20	1.77	36, −58, −11	5.33	66, −52, 19	3.45	63, −7, −17	3.73	48, 17, −41	3.00
10	45, −64, −20	3.58	36, −58, −14	5.18	63, −31, 10	1.94	63, −1, −11	2.02	45, 11, −38	2.13
11	36, −40, −23	1.94	—	—	—	—	63, −4, −17	1.92	—	—
12	—	—	—	—	63, −37, −8	2.31	66, −7, −26	1.77	51, 11, −23	1.68
13	—	—	42, −46, −17	3.93	57, −31, 7	1.81	69, −7, −17	1.83	63, 11, −5	2.36
14	54, −49, −26	1.72	42, −43, −20	3.62	66, −37, −5	3.16	57, −10, −20	2.24	—	—
15	—	—	42, −58, −20	2.91	72, −37, 7	2.85	66, −1, −11	2.14	51, 17, −17	2.02
16	33, −67, −8	1.69	42, −49, −23	2.94	72, −28, 7	2.02	63, −13, −20	2.15	—	—
17	42, −43, −26	2.48	39, −52, −20	5.88	63, −37, 19	1.81	72, −10, −14	1.94	54, 14, −29	1.86
18	—	—	—	—	—	—	57, −10, −23	2.37	—	—
19	—	—	42, −46, −17	2.42	—	—	—	—	—	—
Second level	54, −46, −20	1.93	42, −49, −23	2.84	63, −34, 7	2.81	63, −7, −14	3.01	57, 8, −11	2.61

The peak coordinates and *Z*-statistics are shown for single subject (1–19) and group analyses (second level). Dashes indicate that we could not localize a maximum below a threshold of $p = 0.05$.

The interleaved measurement of 88 axial slices with 1.7 mm thickness (no gap) covered the entire brain, resulting in an isotropic voxel size of 1.7 mm. Additionally, fat saturation was used together with 6/8 partial Fourier imaging and generalized autocalibrating partially parallel acquisitions (Griswold et al., 2002) with an acceleration factor of 2. Seven images without any diffusion weighting (b0-images) were obtained; one before scanning the dMRI sequence and one after each block of 10 diffusion-weighted images. These images were used as anatomical reference for off-line motion correction. Total duration of this scanning session was 15 min.

It has been recommended to use images with high signal-to-noise ratio (SNR) to avoid implausible tracking results (Fillard et al., 2011). We obtained data with high SNR by using parallel acquisition with a 32-channel head coil at a high magnetic field strength (3 tesla), a large number of directions, and seven repetition of the baseline (b = 0) image. This resulted in an SNR of 73 in the white matter of the baseline images and an SNR of 37 in the white matter of the diffusion-weighted (b = 1000) images (DWI). For the averaged b0 image, these SNRs increased to 130 and 83 for the mean of the 60 DWIs. The SNR was measured as mean signal (S) in the white matter divided by the standard deviation (σ) in a background region (free from ghosting or blurring artifacts), i.e., SNR = 0.655 * S/σ. The constant scaling factor (0.655) corrects for the Rician distribution of the background noise (Kaufman et al., 1989; Firbank et al., 1999). For voxelwise analysis of diffusion data, an SNR of 15 in the b0 image has been proposed as sufficient (Smith et al., 2007). Tractography has higher requirements than voxelwise analysis, but the SNR in our data (73), in combination with the high spatial and angular resolution of the measured datasets, should minimize the possibility of finding false-positive connections.

T1. Structural images were acquired with a T1-weighted 3D MP-RAGE with selective water excitation and linear phase encoding. Magnetization preparation consisted of a nonselective inversion pulse. The imaging parameters were TI = 650 ms, TR = 1300 ms, TE = 3.93 ms, $\alpha = 10°$, spatial resolution of 1 mm^3, two averages. To avoid aliasing, oversampling was performed in the read direction (head–foot).

Data analysis

Behavioral

Behavioral data were analyzed with Matlab (version 7.7; MathWorks). Both localizers were matched for task difficulty. We measured, across subjects, 93.36% correct responses for the speech-recognition and 90.98% for the voice-recognition task. There was no significant effect of task (paired *t* test for speech and speaker task; $t_{(18)} = 1.6381$, $p = 0.1188$). The contrast viewing face stimuli versus viewing object stimuli was also matched in difficulty (89.41% correct responses for the visual person category, 90.22% for the visual object category, averaged over the two tasks; paired *t* test for visually presented person and visually presented objects; $t_{(17)} = 0.0138$, $p = 0.4460$).

Functional

Functional data were analyzed with statistical parametric mapping (SPM8; Wellcome Trust Centre for Neuroimaging; http://www.fil.ion.ucl.ac.uk/spm) using standard spatial preprocessing procedures (realignment and unwarp, normalization to MNI standard stereotactic space, and smoothing with an isotropic Gaussian filter, 8 mm at FWHM). Geometric distortions due to susceptibility gradients were corrected by an interpolation procedure based on the B0 map (the field-map). Statistical parametric maps were generated by modeling the evoked hemodynamic response for the different conditions as boxcars convolved with a synthetic hemodynamic response function in the context of the general linear model (Friston et al., 2007). To obtain the individual coordinates of seed and target regions for each subject, the analysis was performed at the single-subject level. Additionally, to compensate for cases where we could not localize seed or target region for a single subject, we localized these regions at the group level. These population-level inferences using BOLD signal changes between conditions of interest were based on a random-effects model that estimated the second-level *t* statistic at each voxel (Friston et al., 2007).

Localizing seed and target regions using functional data

Voice-sensitive areas. For localizing voice-sensitive areas, we used the contrast speaker recognition > speech recognition (Localizer 1; Fig. 2*B*) for each individual subject in MNI space. At the group level, voice-sensitive areas were localized in posterior, middle, and anterior parts of the STS (pSTS, mSTS, and aSTS, respectively) (MNI coordinates: pSTS at 63, −34, 7, $Z = 2.81$; mSTS at 63, −7, −14, $Z = 3.01$; and aSTS at 57, 8, −11, $Z = 2.61$; Table 1). At the individual level, posterior STS could be localized in 14 of 19 subjects, middle STS could be localized in 16 of 19 subjects, and anterior STS could be localized in 11 of 19 subjects.

Face-sensitive area: auditory. The FFA has been found to not only responds to visual stimuli but also to voices during speaker recognition after a brief audiovisual sensory experience (von Kriegstein et al., 2005, 2006, 2008; von Kriegstein and Giraud, 2006). For the contrast speaker recognition > speech recognition in the group analysis, the FFA was located at MNI coordinates 54, −46, −20 ($Z = 1.82$). The FFA could be located in 11 of 19 subjects on the individual level (Table 1). These are

fewer subjects than expected based on similar localizations in previous studies (von Kriegstein et al., 2006, 2008). We attribute this to the relatively short scanning time of the localizer (11 min in contrast to 40 min in previous fMRI designs).

Face-sensitive area: visual. For localizing visual face-sensitive areas, we computed the contrast (person-recognition task + person-matching task) > (mobile-recognition task + mobile-matching task) with visual-only stimuli (Localizer 2; Fig. 2*C*). In the group analysis, the FFA was located at MNI coordinates 42, −49, −23 ($Z = 2.84$). We localized the FFA in 13 of 19 subjects (Table 1).

The coordinates of all three localizers are in line with previous studies (Kanwisher et al., 1997; Belin et al., 2000, 2002; Belin and Zatorre, 2003; von Kriegstein et al., 2003, 2005; von Kriegstein and Giraud, 2004). For subjects in which we were not able to localize the areas of interest (at $p < 0.05$, uncorrected), the coordinates were taken from the group analysis (Table 1).

Localizing seed and target regions for probabilistic tractography in dMRI data. To identify fiber pathways in single-subject dMRI data, we transferred the functionally located coordinates into the individual dMRI space. We moved the transformed localization coordinates to the nearest point of the gray-/white-matter boundary computed from the fractional anisotropy (FA) map (FA > 0.25) and centered a sphere of 5 mm radius on these coordinates (Makuuchi et al., 2009). In addition, we masked seed and target regions with white matter to ensure that we only tracked from white-matter voxels. Due to this masking procedure, we obtained different numbers of voxels in seed and target regions for the individual subjects. To account for these differences, we normalized the tracking results of the individual subjects with respect to the numbers of voxels in seed and target regions (see Connectivity strength, below).

dMRI preprocessing

dMRIs were corrected for participant motion using the seven reference b0 images without diffusion weighting. This was done with linear rigid-body registration (Jenkinson et al., 2002) implemented in FSL (http://www.fmrib.ox.ac.uk/fsl). Motion-correction parameters were interpolated for the 60 diffusion-weighted images and combined with a global registration to the T1 anatomy computed with the same method. The registered images were interpolated to anatomical reference image providing an isotropic voxel resolution of 1.7 mm. The gradient direction for each volume was corrected using the individual rotation parameters. Finally, for each voxel, a diffusion tensor (Basser et al., 1994) was fitted to the data and the FA index (a standard tensor-based measure of tissue anisotropy) was computed (Basser and Pierpaoli, 1996).

Anatomical connectivity was estimated using FDT (FMRIB's Diffusion Toolbox; (http://www.fmrib.ox.ac.uk/fsl/fdt/index.html). The software module BEDPOSTX allowed us to infer a local model of fiber bundles orientations in each voxel of the brain from the measured data (Behrens et al., 2003). We estimated the distribution of up to two crossing fiber bundles in each voxel. The maximal number of two bundles was based on the b-value und the resolution of the data (Behrens et al., 2003, 2007).

For each subject, probabilistic fiber tractography was computed using the software module PROBTRACKX with seed and target masks. This produces an estimate of the most likely location and strength of a pathway between the two areas (Behrens et al., 2007; Johansen-Berg and Behrens, 2009). The connection probability is given by the number of tracts that reach a target voxel from a given seed. We used the standard parameters with 5000 sample tracts per seed voxel, a curvature threshold of 0.2, a step length of 0.5, and a maximum number of steps of 2000. Probabilistic tracking was done between the face-sensitive area in the fusiform gyrus and voice-recognition areas in the right STS (FFA–pSTS, FFA–mSTS, FFA—aSTS; and between pSTS–mSTS, mSTS–aSTS, pSTS–aSTS). This was done for both identified FFA coordinates. We will refer to the FFA localized by the auditory localizer as crossmodal FFA (cFFA) and the FFA coordinate localized by the visual experiment as visual FFA (vFFA).

Connectivity strength

The connectivity strength was determined from the number of tracts from each seed that reached the target (Eickhoff et al., 2010; Forstmann et al., 2010). We tracked in both directions for each pair of seed and target region and summed the resulting two connectivity measures to obtain one connectivity measure per pair of regions. The obtained measure relates to the probability of a connection between both areas (Bridge et al., 2008).

Note that statistical thresholding of probabilistic tractography is an unsolved statistical issue (Morris et al., 2008). For the binary decision of whether a specific connection exists, we considered the connectivity between two brain areas as reliable if at least 10 pathways between each pair of seed and target masks (sum of both directions) were present (Makuuchi et al., 2009). With this fixed arbitrary threshold we aimed at both reducing false-positive connections and staying sensitive enough to not miss true connections (Heiervang et al., 2006; Johansen-Berg et al., 2007). In probabilistic tractography, it is possible that a specific connection cannot be found in all subjects due to variations in gyrification and other anatomical factors. In particular, this might be the case for connections with a high curvature along the tract, which are most challenging for current tractography algorithms. Therefore we assume that, if the connection can be repeatedly found with this conservative threshold in a number of participants (≥50%), the probability is high that this connection exists in all subjects (Saur et al., 2008; Makuuchi et al., 2009; Doron et al., 2010). After thresholding, we therefore counted the number of subjects who showed a connection for the specific pair and normalized by the number of all participants ($n = 19$).

Additionally, as a quantitative measure of connectivity between each pair of seed and target region in a single subject, we calculated connectivity indices (Makuuchi et al., 2009; Eickhoff et al., 2010). For all subjects, the index was defined by counting the number of connected pathways between each pair of seed and target masks (in both directions) and dividing it by the number of all connection pathways for all seed and target masks of the individual subject (Eickhoff et al., 2010). With this normalization we accounted for potential large interindividual differences in connectivity strength between subjects. To account for difference in size of target and seed regions, we subdivided this index by the number of voxels in seed and target regions (see Fig. 4) (Makuuchi et al., 2009; Eickhoff et al., 2010). This calculation and normalization of the connectivity indices was done to quantify the structural connectivity between face-sensitive area in the fusiform gyrus (located with two different contrasts) and voice-recognition areas in the right STS (FFA–pSTS, FFA–mSTS, FFA–aSTS and between pSTS–mSTS, mSTS–aSTS, pSTS–aSTS). As the connectivity indices were not normally distributed, we used nonparametric tests to test significance of differences between these indices.

Distance correction

For computing the distance along the connecting pathway between seed and target regions, we first created two probabilistic connectivity maps (with and without distance correction) and divided these maps to compute a map of average pathway length. We masked this distance map with the corresponding seed and target regions and extracted the relevant distance values of the regions of interest for each subject. We averaged across and within each pair of seed and target region and thereby obtained one distance value for each pair of seed and target region.

At the group level, we tested for potential differences in length of the pathways between the different seed and target pairs. Subjects without a connection in either of the two pairs of seed and target region within one comparison were excluded from this analysis. The pathway between FFA and anterior STS was longer than the pathway between FFA and posterior STS (paired *t* test: 86.35 vs 92.84, $t_{(29)} = 2.3613$, $p = 0.0251$). The pathway between FFA and anterior STS was also longer than the pathway between FFA and middle STS (paired *t* test: 96.58 vs 89.67, $t_{(35)} = 2.4295$, $p = 0.0204$). In contrast, there was no significant difference in pathway length between FFA and posterior STS compared with the pathway length between FFA and middle STS (paired *t* test: 88.94 vs 91.35, $t_{(30)} = 0.7110$, $p = 0.4826$). The lengths of the connecting pathways from cFFA and vFFA to the regions in STS also did not differ significantly (paired *t* test: 90.4182 vs 94.4200, $t_{(48)} = 1.8312$, $p = 0.0733$).

The comparison of the path lengths within the voice-sensitive regions in the STS showed that the pathway from the posterior STS to the ante-

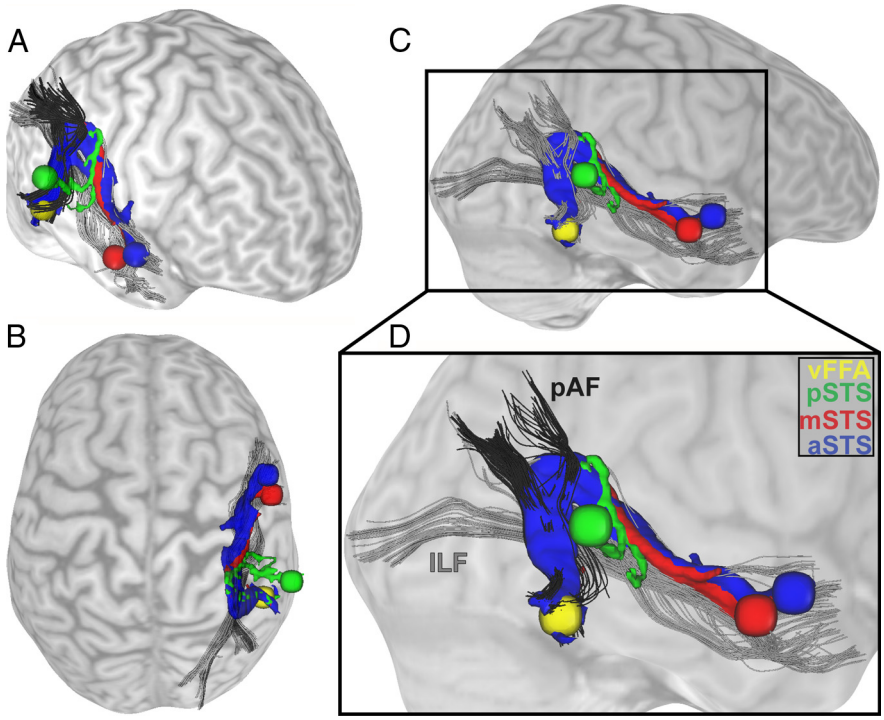

Figure 3. Structural connections between voice- and face-sensitive areas as found with probabilistic fiber tracking. Results of a single, representative participant's dMRI are shown as connectivity distribution. Probabilistic tractography was done in both directions: from the visual FFA as seed region to the three target regions in the STS and vice versa from the three seed regions in the STS to the FFA as target region. Seed and target regions were localized with a functional localizer and are displayed as spheres with a radius of 5 mm (yellow, FFA; blue, anterior part of the STS; red, middle part of the STS; green, posterior part of the STS). The structural connections between FFA and STS are colored correspondingly to their seed and target regions in the STS. The connections are displayed from different views: right frontal (***A***), top (***B***), and right side (***C***), plus a detailed view (***D***). As anatomical landmarks, the inferior longitudinal fasciculus (ILF) is shown in gray (***B–D***) and the posterior part of the arcuate fasciculus (pAF) is shown in black (***C***, ***D***) (Catani et al., 2003, 2005; Catani and Thiebaut de Schotten, 2008). Close to the FFA, the connecting pathways follow the posterior part of the arcuate fasciculus before taking a turn to follow the inferior longitudinal fasciculus and reach the STS regions.

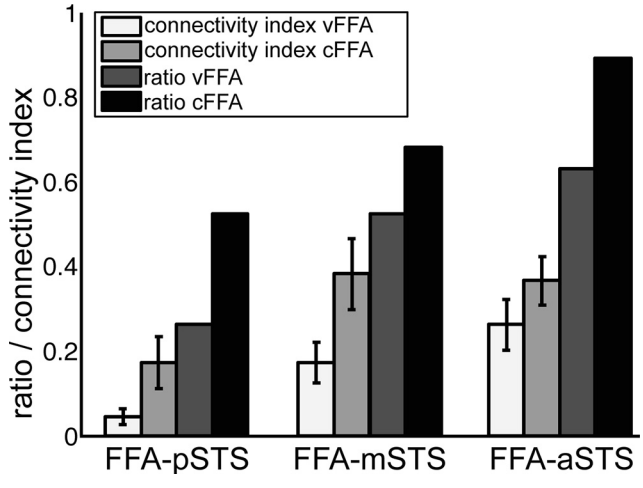

Figure 4. Quantitative analyses of structural connections between FFA and STS regions. Seed and target regions were the face-sensitive region (vFFA, cFFA) and the voice-sensitive regions (posterior, middle, and anterior STS). Structural connectivity indices (white and light gray) were calculated for each connection in individuals ($n = 19$ participants; error bars show SEM). Structural connectivity index was defined as the number of connected voxels divided by the overall number of connected voxels per participant and by the number of voxels of seed and target regions (Eickhoff et al., 2010). The structural connectivity index was multiplied by 100 for a common display together with ratio values: The ratio (dark gray) shows the number of subjects who showed a connection normalized by the number of all participants ($n = 19$).

rior STS was longer than to the middle STS (paired t test: 65.56 vs 52.05, $t_{(17)} = 2.4007$, $p = 0.0281$). Additionally, the pathway from the anterior STS to the posterior STS was longer than to the middle STS (paired t test: 65.56 vs 43.05, $t_{(17)} = 4.3187$, $p = 0.0005$). The length of the pathways between middle and posterior STS and middle and anterior STS did not differ significantly (paired t test: 55.79 vs 45.47, $t_{(18)} = 1.9850$, $p = 0.0626$).

Since in FSL the number of tracts from seed to target is not automatically corrected for distance and because we wanted to ensure that our results were not caused by any distance bias toward shorter connections, we multiplied the individual connectivity indices with the distance value of each seed–target connection. This approach made it less probable that physical distance alone caused our results (Tomassini et al., 2007).

Results

Structural connectivity between face- and voice-recognition areas

The results show evidence that face- and voice-sensitive areas in the human brain are connected via direct structural connections (Figs. 3–5).

We used probabilistic tractography to investigate the structural connections between the functionally localized FFA and voice-sensitive areas in STS. We found evidence for structural connections between FFA and aSTS (vFFA in 12 of 19 subjects; cFFA in 17 of 19 subjects), between FFA and mSTS (vFFA in 10 of 19 subjects; cFFA in 13 of 19 subjects), and between FFA and pSTS (cFFA 10 of 19 subjects; Fig. 4). The structural connection between vFFA and pSTS was present only in a few subjects (vFFA 5 of 19 subjects).

Testing differences between connection indices revealed that the FFA (independent of the type of localization) is significantly stronger connected with middle and anterior STS than with posterior STS (nonparametric Friedman's test: mean ranks of FFA–pSTS, FFA–mSTS, and FFA–aSTS were 1.43, 2.12, and 2.45, respectively; $X^2_{(2)} = 21.89$; $p < 0.0001$ and *post hoc* Wilcoxon signed ranks tests: FFA–aSTS vs FFA–pSTS: $Z = 3.77$, $p < 0.0001$; FFA–mSTS vs FFA–pSTS: $Z = 2.81$, $p = 0.0049$, both Bonferroni corrected; Fig. 4). In contrast, there is no significant difference in connectivity strength from FFA to aSTS and mSTS (Wilcoxon signed ranks test, $Z = 1.04$, $p = 0.2961$), which may indicate that FFA is equally strong connected to anterior and middle STS.

Furthermore, FFA localized by the contrast "speaker recognition versus speech recognition" (cFFA) is significantly more strongly connected with the STS regions than the FFA localized by the contrast "faces versus objects" (vFFA) (Wilcoxon signed rank test, cFFA vs vFFA, $Z = 2.96$, $p = 0.0031$, Fig. 4).

We obtained qualitatively the same results with distance-corrected connectivity indices.

Structural connectivity between voice-sensitive areas

We also investigated the structural connectivity between the voice-sensitive regions within the STS. The probabilistic tracking indicated

that posterior, middle, and anterior parts of the STS are structurally connected. We found evidence for structural connections between pSTS–aSTS in 15 of 19 subjects, and between mSTS–aSTS and between pSTS–mSTS in all 19 subjects. Comparing the connectivity indices between the STS regions showed that there is no significant difference in connections within the STS regions (nonparametric Friedman's test: mean ranks of pSTS–aSTS, pSTS–mSTS, and aSTS–mSTS were 1.55, 2.18, and 2.26, respectively; $X^2_{(2)} = 5.84$; $p = 0.0539$). We obtained qualitatively the same results with distance-corrected connectivity indices.

Discussion

We found evidence for direct structural connections between face- and voice-recognition areas in the human brain by combined functional magnetic resonance and diffusion-weighted imaging. Our findings show that the FFA has significantly larger structural connectivity to middle and anterior than to posterior areas of the voice-sensitive STS. Additionally, we provide evidence that the three different voice-sensitive regions within the STS are all connected with each other.

The direct link between FFA and the voice-sensitive STS indicates that person identity processing in the human brain is not only based on integration at a supramodal stage as shown in Figure 1*A*. Rather, face- and voice-sensitive areas can exchange information directly with each other (Fig. 1*B*).

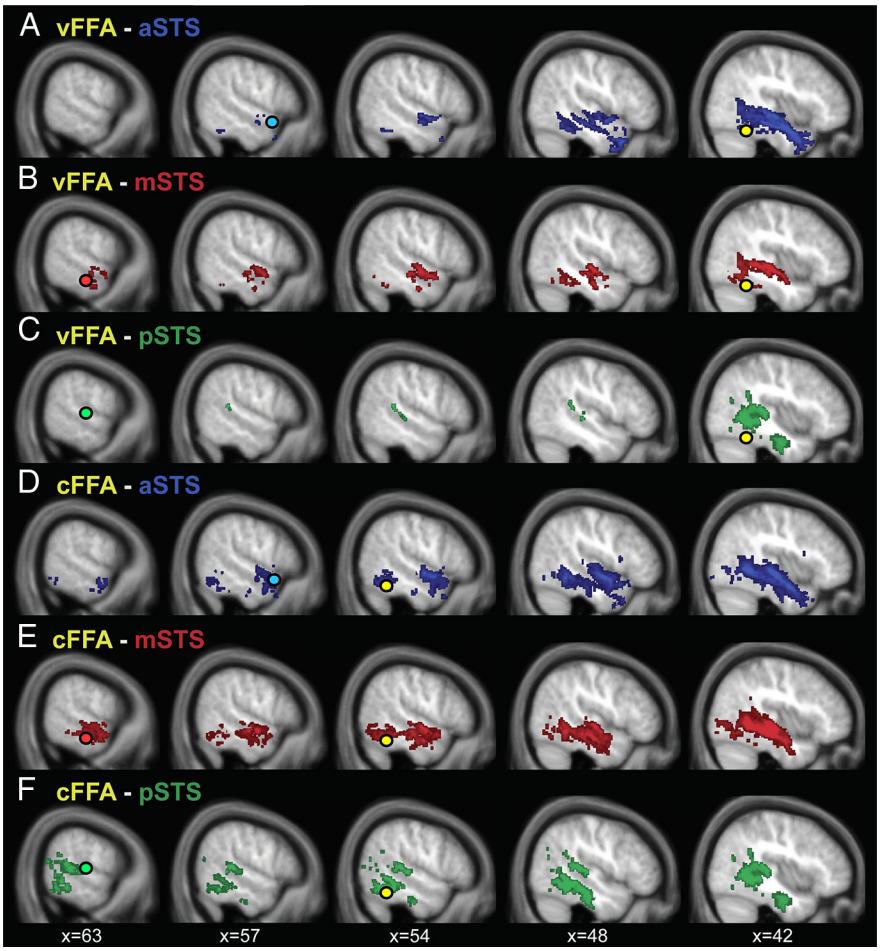

Figure 5. Group overlay of probabilistic pathways between voice-sensitive STS and face-sensitive FFA. Connectivity distributions of 19 participants' dMRI data were binarized, thresholded at 10 paths per voxel at the individual subject level, and overlaid for display purposes. There are connections between the FFA (yellow circle) and the anterior part of the STS (blue circle; ***A***, ***D***), the middle part of the STS (red circle; ***B***, ***E***), and the posterior part of the STS (green circle, ***C***, ***F***). The connectivity distributions are colored correspondingly to the STS seed and target masks. ***A–C***, Tracking results for FFA localized with the visual localizer. ***D–F***, FFA localized with the auditory localizer. Tracking results are depicted on the averaged T1 scan of the 19 subjects.

Structural connectivity between FFA and middle/anterior voice-sensitive STS fits well with previous findings, which reported functional connectivity of the FFA to the same middle/anterior voice-sensitive STS region during voice-recognition tasks (von Kriegstein et al., 2005; von Kriegstein and Giraud, 2006). The results of our study also integrate well with recent developments in multisensory research showing that information from different modalities interact earlier and on lower processing levels than traditionally thought (Cappe et al., 2010; Kayser et al., 2010; Klinge et al., 2010; Beer et al., 2011; for review see Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008). We assume that direct connections between FFA and voice-sensitive cortices are especially relevant in the context of person identification. For other aspects of face-to-face communication, such as speech or emotion recognition, other connections might be more relevant. For example, speech recognition may benefit from the integration of fast-varying dynamic visual and auditory information (Sumby and Pollack, 1954). In this case, direct connections between visual movement areas and auditory cortices might be used (Ghazanfar et al., 2008; von Kriegstein et al., 2008; Arnal et al., 2009). Additionally, interaction mechanisms that integrate basic auditory and visual stimuli (Noesselt et al., 2007) might also be involved in voice and face integration.

The structural connections between FFA and middle/anterior voice-sensitive STS overlap with major white-matter pathways, the inferior longitudinal fasciculus, and the posterior part of the arcuate fasciculus (Fig. 3) (Catani et al., 2003, 2005; Catani and Thiebaut de Schotten, 2008). The posterior segment of the arcuate fasciculus connects the inferior parietal lobe to posterior temporal areas. The inferior longitudinal fasciculus connects occipital and temporal lobes via a ventral tract and is involved in face processing (Fox et al., 2008; Thomas et al., 2009). Our results suggest a connection between both pathways including fibers running in dorsal direction from the FFA parallel to the arcuate fasciculus and the lateral boundary of the ventricles and bending anterior to follow the inferior longitudinal fasciculus to connect with the voice-sensitive STS regions.

In the present study, we aimed at specificity of the structural connectivity findings and defined target and seed regions functionally for each individual subject. It has been shown that FFA responds to both facial features and face identity (Kanwisher et al., 1997; Eger et al., 2004; Rotshtein et al., 2005; Dachille and James, 2010) and probably not all of FFA analyzes features that are relevant for exchange with voice-sensitive areas for identity recognition. We therefore used two different contrasts to localize the FFA (Fig. 2). First, we used a conventional visual localizer contrasting conditions containing faces with conditions containing objects. This contrast has the advantage that it is the standard

contrast to localize the FFA, but the disadvantage that it is not particularly geared toward face identity processing. In contrast, the second localizer emphasizes identity recognition (speaker recognition after face learning > speech recognition after face learning) (von Kriegstein and Giraud, 2006; von Kriegstein et al., 2008). We found evidence that both identified FFA coordinates (vFFA and cFFA) were structurally connected to the voice-sensitive areas in the STS. However, the coordinate of the more specific localizer (cFFA) was more strongly connected to voice-sensitive STS than the coordinate of the standard localizer (vFFA). We speculate that this specificity of structural connections between FFA and voice-sensitive STS plays a functional role in person recognition.

Structural connectivity seems to exist between the FFA and all voice-sensitive regions in the STS, but appears to be particularly strong for anterior/middle STS. This is intriguing because our analyses of pathway lengths show that anterior STS is further away from the FFA than posterior STS. This connectivity validates our results since tracking artifacts would rather show stronger connectivity for regions in close proximity to each other (Anwander et al., 2007; Tomassini et al., 2007; Bridge et al., 2008; Eickhoff et al., 2010). Our findings suggest that areas that are involved in voice identity processing show particularly strong connections to the FFA. Conversely, posterior voice-sensitive regions in STS have been found to process acoustic parameters of voices, which suggests limited need for exchange of information between posterior STS and FFA (Belin and Zatorre, 2003; von Kriegstein et al., 2003, 2007; von Kriegstein and Giraud, 2004; Andics et al., 2010).

Our results also imply that posterior, middle, and anterior STS are structurally connected with each other. This was expected and also complements previous functional connectivity findings that showed functional connectivity between voice-sensitive STS regions during speaker recognition (von Kriegstein and Giraud, 2004).

Tractography is the only method currently available to investigate anatomical information in terms of fiber bundles in humans *in vivo* (Conturo et al., 1999; Dyrby et al., 2007). It is still a new method, but some of its initial limitations have been recently surmounted by new analysis techniques (Johansen-Berg and Behrens, 2006; Johansen-Berg and Rushworth, 2009; Jones, 2010; Jones and Cercignani, 2010; Chung et al., 2011). For example, tracking close to gray matter usually results in limited connectivity findings (Anwander et al., 2007). We addressed this issue by only tracking between voxels with certain white-matter strength (FA > 0.25). Another limitation is that deterministic fiber tracking follows only the main diffusion direction of each voxel, which can result in poor connectivity reconstruction (Johansen-Berg and Behrens, 2009). We therefore used probabilistic tracking. Probabilistic tracking takes computed distributions of possible fiber directions of the pathway into account, which renders tracking more robust to noise and enables the detection of crossing fibers (Behrens et al., 2007). It also has the advantage of providing a quantitative measure of connectivity strength. Despite these methodological advances, probabilistic tracking algorithms are, in principle, not capable of proving the existence of a connection between any two regions and can only suggest potential connections (Morris et al., 2008). They might also miss or suggest false-positive pathways (Fillard et al., 2011). A recently developed statistical method that compares structural connections with a random pattern of connectivity to determine significance might provide a framework to address these issues in the future (Morris et al., 2008).

Direct connections between auditory and visual person-processing areas suggest that the assessment of person-specific information does not necessarily have to be mediated by supramodal cortical structures (like so-called modality-free person identity nodes; Fig. 1*A*) (Bruce and Young, 1986; Burton et al., 1990; Ellis et al., 1997). It could also result from direct cross-modal interactions between voice- and face-sensitive regions (von Kriegstein et al., 2005; von Kriegstein and Giraud, 2006). Direct structural connections between FFA and STS voice regions are a prerequisite for the model in Figure 1*B* (von Kriegstein and Giraud, 2006; von Kriegstein et al., 2008). This model has been formulated in the framework of a predictive coding account of brain function (Rao and Ballard, 1999; Friston, 2005; Kiebel et al., 2008). In this view, direct reciprocal interactions between auditory and visual sensory-processing steps serve to exchange predictive (i.e., constraining) information about the person's characteristics. These predictive signals can be used to constrain possible interpretation of unisensory, noisy, or ambiguous sensory input and thereby optimize recognition (Ernst and Banks, 2002).

The model (Fig. 1*B*) was developed based on behavioral and fMRI findings. In auditory-only conditions, voices are better recognized when subjects have had brief audiovisual training with a video of the respective speakers (in contrast to matched-control training) (Sheffert and Olson, 2004; von Kriegstein and Giraud, 2006; von Kriegstein et al., 2008). This is at odds with the conventional model (Fig. 1*A*) because it implies that face information can be used for voice recognition even in the absence of visual input. Furthermore, fMRI studies revealed an involvement of the FFA in auditory-only voice recognition (von Kriegstein et al., 2005, 2006, 2008; von Kriegstein and Giraud, 2006). Studies on prosopagnosic and normal subjects have shown that the FFA is relevant for optimal voice-recognition performance (von Kriegstein et al., 2008). FFA involvement during voice recognition could be reconciled with the conventional model if it allowed the transfer of information from the auditory modality to the FFA via a supramodal stage. However, this is at odds with functional connectivity findings which show that face-sensitive (FFA) and voice-sensitive areas (in the STS) are functionally connected during auditory-only speaker recognition, while the functional connectivity to supramodal regions plays a minor role (von Kriegstein et al., 2005, 2006; von Kriegstein and Giraud, 2006). Together with this previous work, our findings suggest that the identified structural connection pattern between FFA and voice-sensitive areas in the STS serves optimized voice recognition in the human brain.

In summary, our findings imply that conventional models of person recognition need to be modified to take a direct exchange of information between auditory and visual person-recognition areas into account.

References

Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánszky Z (2010) Neural mechanisms for voice recognition. Neuroimage 52:1528–1540.

Anwander A, Tittgemeyer M, von Cramon DY, Friederici AD, Knösche TR (2007) Connectivity-based parcellation of Broca's area. Cereb Cortex 17:816–825.

Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29:13445–13453.

Basser PJ, Pierpaoli C (1996) Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. J Magn Reson B 111:209–219.

Basser PJ, Mattiello J, LeBihan D (1994) Estimation of the effective self-diffusion tensor from the NMR spin-echo. J Magn Reson B 103:247–254.

Beer AL, Plank T, Greenlee MW (2011) Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. Exp Brain Res. Advance online publication. Retrieved May 15, 2011. doi:10.1007/s00221-011-2715-y.

Behrens TE, Woolrich MW, Jenkinson M, Johansen-Berg H, Nunes RG, Clare S, Matthews PM, Brady JM, Smith SM (2003) Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magn Reson Med 50:1077–1088.

Behrens TE, Berg HJ, Jbabdi S, Rushworth MF, Woolrich MW (2007) Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? Neuroimage 34:144–155.

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

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

Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13:17–26.

Berman MG, Park J, Gonzalez R, Polk TA, Gehrke A, Knaffla S, Jonides J (2010) Evaluating functional localizers: the case of the FFA. Neuroimage 50:56–71.

Bridge H, Thomas O, Jbabdi S, Cowey A (2008) Changes in connectivity after visual cortical brain damage underlie altered visual function. Brain 131:1433–1444.

Bruce V, Young A (1986) Understanding face recognition. Br J Psychol 77:305–327.

Burton AM, Bruce V, Johnston RA (1990) Understanding face recognition with an interactive activation model. Br J Psychol 81:361–380.

Cappe C, Thut G, Romei V, Murray MM (2010) Auditory-visual multisensory interactions in humans: timing, topography, directionality, and sources. J Neurosci 30:12572–12580.

Catani M, Thiebaut de Schotten M (2008) A diffusion tensor imaging tractography atlas for virtual in vivo dissections. Cortex 44:1105–1132.

Catani M, Jones DK, Donato R, Ffytche DH (2003) Occipito-temporal connections in the human brain. Brain 126:2093–2107.

Catani M, Jones DK, ffytche DH (2005) Perisylvian language networks of the human brain. Ann Neurol 57:8–16.

Chung HW, Chou MC, Chen CY (2011) Principles and limitations of computational algorithms in clinical diffusion tensor MR tractography. AJNR 32:3–13.

Conturo TE, Lori NF, Cull TS, Akbudak E, Snyder AZ, Shimony JS, McKinstry RC, Burton H, Raichle ME (1999) Tracking neuronal fiber pathways in the living human brain. Proc Natl Acad Sci U S A 96: 10422–10427.

Dachille L, James T (2010) The role of isolated face features and feature combinations in the fusiform face area. J Vis 10:660.

Doron KW, Funk CM, Glickstein M (2010) Fronto-cerebellar circuits and eye movement control: A diffusion imaging tractography study of human cortico-pontine projections. Brain Res 1307:63–71.

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. Neuron 57:11–23.

Dyrby TB, Søgaard LV, Parker GJ, Alexander DC, Lind NM, Baaré WF, Hay-Schmidt A, Eriksen N, Pakkenberg B, Paulson OB, Jelsing J (2007) Validation of in vitro probabilistic tractography. Neuroimage 37:1267–1277.

Eger E, Schyns PG, Kleinschmidt A (2004) Scale invariant adaptation in fusiform face-responsive regions. Neuroimage 22:232–242.

Eickhoff SB, Jbabdi S, Caspers S, Laird AR, Fox PT, Zilles K, Behrens TE (2010) Anatomical and functional connectivity of cytoarchitectonic areas within the human parietal operculum. J Neurosci 30:6409–6421.

Ellis HD, Jones DM, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. Br J Psychol 88:143–156.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429–433.

Fillard P, Descoteaux M, Goh A, Gouttard S, Jeurissen B, Malcolm J, Ramirez-Manzanares A, Reisert M, Sakaie K, Tensaouti F, Yo T, Mangin JF, Poupon C (2011) Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. Neuroimage 56:220–234.

Firbank MJ, Coulthard A, Harrison RM, Williams ED (1999) A comparison of two methods for measuring the signal to noise ratio on MR images. Phys Med Biol 44:N261–N264.

Föcker J, Hölig C, Best A, Röder B (2011) Crossmodal interaction of facial and vocal person identity information: an event-related potential study. Brain Res 1385:229–245.

Forstmann BU, Anwander A, Schäfer A, Neumann J, Brown S, Wagenmakers EJ, Bogacz R, Turner R (2010) Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. Proc Natl Acad Sci U S A 107:15916–15920.

Fox CJ, Iaria G, Barton JJ (2008) Disconnection in prosopagnosia and face processing. Cortex 44:996–1009.

Fox CJ, Iaria G, Barton JJ (2009) Defining the face processing network: optimization of the functional localizer in fMRI. Hum Brain Mapp 30:1637–1651.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond Biol 360:815–836.

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138.

Friston K, Ashburner J, Kiebel S, Nichols T, Penny W (2007) Statistical parametric mapping: the analysis of functional brain images. New York: Academic.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278–285.

Ghazanfar AA, Chandrasekaran C, Logothetis NK (2008) Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J Neurosci 28:4457–4469.

Griswold MA, Jakob PM, Heidemann RM, Nittka M, Jellus V, Wang J, Kiefer B, Haase A (2002) Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 47:1202–1210.

Heiervang E, Behrens TE, Mackay CE, Robson MD, Johansen-Berg H (2006) Between session reproducibility and between subject variability of diffusion MR and tractography measures. Neuroimage 33:867–877.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Johansen-Berg H, Behrens TE (2006) Just pretty pictures? What diffusion tractography can add in clinical neuroscience. Curr Opin Neurol 19:379–385.

Johansen-Berg H, Behrens TE (2009) Diffusion MRI: from quantitative measurement to in vivo neuroanatomy. New York: Academic.

Johansen-Berg H, Rushworth MF (2009) Using diffusion imaging to study human connectional anatomy. Annu Rev Neurosci 32:75–94.

Johansen-Berg H, Della-Maggiore V, Behrens TE, Smith SM, Paus T (2007) Integrity of white matter in the corpus callosum correlates with bimanual co-ordination skills. Neuroimage 36:T16–T21.

Jones DK (2010) Challenges and limitations of quantifying brain connectivity in vivo with diffusion MRI. Imaging Med 2:341–355.

Jones DK, Cercignani M (2010) Twenty-five pitfalls in the analysis of diffusion MRI data. NMR Biomed 23:803–820.

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.

Kaufman L, Kramer DM, Crooks LE, Ortendahl DA (1989) Measuring signal-to-noise ratios in MR imaging. Radiology 173:265–267.

Kayser C, Logothetis NK (2007) Do early sensory cortices integrate cross-modal information? Brain Struct Funct 212:121–132.

Kayser C, Logothetis NK, Panzeri S (2010) Visual enhancement of the information representation in auditory cortex. Curr Biol 20:19–24.

Kiebel SJ, Daunizeau J, Friston KJ (2008) A hierarchy of time-scales and the brain. PLoS Comput Biol 4:e1000209.

Klinge C, Eippert F, Röder B, Büchel C (2010) Corticocortical connections mediate primary visual cortex responses to auditory stimulation in the blind. J Neurosci 30:12798–12805.

Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A 104:20600–20605.

Makuuchi M, Bahlmann J, Anwander A, Friederici AD (2009) Segregating the core computational faculty of human language from working memory. Proc Natl Acad Sci U S A 106:8362–8367.

Morris DM, Embleton KV, Parker GJ (2008) Probabilistic fibre tracking: differentiation of connections from chance events. Neuroimage 42:1329–1339.

Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze HJ, Driver J (2007) Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J Neurosci 27:11431–11441.

Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9:97–113.

Rajimehr R, Young JC, Tootell RB (2009) An anterior temporal face patch in human cortex, predicted by macaque maps. Proc Natl Acad Sci U S A 106:1995–2000.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Reese TG, Heid O, Weisskoff RM, Wedeen VJ (2003) Reduction of eddy-current-induced distortion in diffusion MRI using a twice-refocused spin echo. Magn Reson Med 49:177–182.

Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat Neurosci 8:107–113.

Saur D, Kreher BW, Schnell S, Kümmerer D, Kellmeyer P, Vry MS, Umarova R, Musso M, Glauche V, Abel S, Huber W, Rijntjes M, Hennig J, Weiller C (2008) Ventral and dorsal pathways for language. Proc Natl Acad Sci U S A 105:18035–18040.

Sergent J, Ohta S, MacDonald B (1992) Functional neuroanatomy of face and object processing: a positron emission tomography study. Brain 115:15–36.

Sheffert SM, Olson E (2004) Audiovisual speech facilitates voice learning. Percept Psychophys 66:352–362.

Smith SM, Johansen-Berg H, Jenkinson M, Rueckert D, Nichols TE, Miller KL, Robson MD, Jones DK, Klein JC, Bartsch AJ, Behrens TE (2007) Acquisition and voxelwise analysis of multi-subject diffusion data with tract-based spatial statistics. Nat Protoc 2:499–503.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215.

Thomas C, Avidan G, Humphreys K, Jung KJ, Gao F, Behrmann M (2009) Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nat Neurosci 12:29–31.

Tomassini V, Jbabdi S, Klein JC, Behrens TE, Pozzilli C, Matthews PM, Rushworth MF, Johansen-Berg H (2007) Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. J Neurosci 27:10259–10269.

von Kriegstein K, Giraud AL (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22:948–955.

von Kriegstein K, Giraud AL (2006) Implicit multisensory associations influence voice recognition. PLoS Biol 4:e326.

von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17:48–55.

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud AL (2005) Interaction of face and voice areas during speaker recognition. J Cogn Neurosci 17:367–376.

von Kriegstein K, Kleinschmidt A, Giraud AL (2006) Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. Cereb Cortex 16:1314–1322.

von Kriegstein K, Smith DR, Patterson RD, Ives DT, Griffiths TD (2007) Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr Biol 17:1123–1128.

von Kriegstein K, Dogan O, Grüter M, Giraud AL, Kell CA, Grüter T, Kleinschmidt A, Kiebel SJ (2008) Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A 105:6747–6752.