Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus

Thomas Ethofer a,b,* , Johannes Bretscher a, Sarah Wiethoff c, Jeanne Bisch a, Sarah Schlipf a, Dirk Wildgruber a, Benjamin Kreifelts a

a Department of General Psychiatry, University of Tübingen, Tübingen, Germany
b Department of Biomedical Resonance, University of Tübingen, Tübingen, Germany
c Department of Neurology, University of Tübingen, Tübingen, Germany

A R T I C L E   I N F O

Article history:
Accepted 26 February 2013
Available online 16 March 2013

Keywords:
Superior temporal sulcus
Orbitofrontal cortex
Voice
Face
Audiovisual integration
Social signals

A B S T R A C T

It was the aim of this study to delineate the areas along the right superior temporal sulcus (STS) for processing of faces, voices, and face–voice integration using established functional magnetic resonance imaging (fMRI) localizers and to assess their structural connectivity profile with diffusion tensor imaging (DTI). We combined this approach with an fMRI adaptation design during which the participants judged emotions in facial expressions and prosody and demonstrated response habituation in the orbitofrontal cortex (OFC) which occurred irrespective of the sensory modality. These functional data were in line with DTI findings showing separable fiber projections of the three different STS modules converging in the OFC which run through the external capsule for the voice area, through the dorsal superior longitudinal fasciculus (SLF) for the face area and through the ventral SLF for the audiovisual integration area. The OFC was structurally connected with the supplementary motor area (SMA) and activation in these two areas was correlated with faster stimulus evaluation during repetition priming. Based on these structural and functional properties, we propose that the OFC is part of the extended system for perception of emotional information in faces and voices and constitutes a neural interface linking sensory areas with brain regions implicated in generation of behavioral responses.

© 2013 Elsevier Inc. All rights reserved.

Introduction

Faces and voices carry a wealth of socially relevant cues including age, gender, identity, attractiveness, as well as intentions, attitudes and the current affective state of the protagonists. Individuals living in complex social systems need to extract this information in a fast and efficient manner. Consequently, the majority of humans can be regarded as face and voice experts. To predict the emotional state, intentions, attitudes, and future actions of other people, correct interpretation of facial and vocal features that can dynamically change during natural communication, such as eye gaze position, facial expression, and speech melody (prosody), is required.

Current models on face perception (Haxby et al., 2000; Ishai, 2008; Tsao and Livingstone, 2008) propose a functional specialization of the face-sensitive cortices along the posterior superior temporal sulcus (pSTS, Puce et al., 1998) for processing of such dynamic facial cues while the occipital face area (e.g. Halgren et al., 1999) and the fusiform face area (Kanwisher et al., 1997) have been proposed as critical sites for assessment of invariant facial features (e.g., identity, gender). This model is supported by neurophysiological data obtained in non-human primates (Hasselmo et al., 1989; Perrett et al., 1985, 1992), observations made in patients with brain lesions (Akiyama et al., 2006; Campbell et al., 1990; Grüsser and Landis, 1991), and neuroimaging experiments comparing top-down (judgment of identity versus eye gaze, Hoffman and Haxby, 2000) or bottom-up effects (presentation of dynamic versus static stimuli, Pitcher et al., 2011) during face processing.

The neural correlates for voice perception are less well understood than those for processing of faces (Latinus and Belin, 2011). However, voice-sensitive areas have been described along the superior temporal sulcus (STS, Belin et al., 2000). In analogy to the face processing system, separate pathways for assessment of invariant and changeable vocal features have been identified by functional magnetic resonance imaging (fMRI) studies (for a detailed review see Campanella and Belin, 2007). While invariant information expressed in the voice (e.g., gender, identity) has been shown to be represented in the anterior STS adjacent to the temporal pole (aSTS, Belin and Zatorre, 2003; Charest et al., 2013; Latinus et al., 2011; von Kriegstein et al., 2003; von Kriegstein and Giraud, 2006), a modulation of the response amplitude by emotional prosody has been demonstrated for the middle part of the STS (mSTS, Ethofer et al., 2006b; Grandjean et al., 2005) which has been shown to be particularly sensitive to specific acoustic properties (Andics et al., 2010; Kriegstein and Giraud, 2004; Wiethoff et al., 2009).
et al., 2008). In agreement with these results, the spatial activation pattern of voice-sensitive areas along the STS can be employed to successfully predict which emotion (Ethofer et al., 2009b) or which speaker (Formisano et al., 2008) was perceived by the listener.

While integration of auditory and visual speech-related signals is thought to rely on the left STS (Aarnal et al., 2009; Blank and von Kriegstein, 2012), integration of socially relevant signals from emotional prosody and facial expressions has been shown to occur at the overlap of right hemispheric face- and voice-sensitive STS cortices (Kreifelts et al., 2009; Szyck et al., 2008; Wright et al., 2003). Neuroimaging studies investigating effective connectivity (Friston et al., 1997, 2003) of these STS regions revealed interactions with the FFA (Kreifelts et al., 2007; Muller et al., 2012) and the amygdala (Muller et al., 2012) during audiovisual integration of emotional signals. A recent neuroimaging study demonstrated direct connections between STS areas responsive to voice identity and the FFA (Blank et al., 2011). Apart from that, however, it is unknown which other brain areas are structurally interconnected with the different parts of the STS. Previous lesion studies (Hornak et al., 1996, 2003) demonstrated impaired recognition of both emotional facial expressions and prosody in patients with damaged orbitofrontal cortex (OFC), an area which is long known from animal studies to receive input from multiple senses (Jones and Powell, 1970) and thus fulfills the criteria of a multisensory convergence zone (Driver and Noesselt, 2008; Mesulam, 1998). In line with these findings, neuroimaging data indicated enhanced activation during active judgment of emotions in faces and voices (e.g. Ethofer et al., 2006a; Sabatini et al., 2011). Based on these convergent results, we predicted structural fiber connections towards and functional activation within the orbital part of the inferior frontal cortex during processing of social signals in faces and voices. To directly test this hypothesis, we combined diffusion tensor imaging (DTI) with a factorial adaptation fMRI paradigm which was specifically designed to test for regional habituation during repeated exposure to faces, voices, and face–voice combinations. Such attenuation of brain responses has been termed repetition suppression — a robust phenomenon which occurs consistently after repeated presentation of identical stimuli and can be exploited to examine which brain regions participate in processing of a certain stimulus type. It has been proposed that repetition suppression reflects top-down mediated perceptual expectations (Summerfield et al., 2008) as well as bottom-up sharpening of neural responses (Larsson and Smith, 2012) which typically results in more accurate and faster behavioral responses (Grill-Spector et al., 2006). Therefore, adaptation designs additionally offer the opportunity to reveal the neural structures which mediate repetition priming effects (Schacter and Buckner, 1998). It should be noted, however, that repetition suppression has also been found outside of neural systems engaged in processing of a particular stimulus type presumably via carry-over effects from other brain areas (Mur et al., 2010) questioning the specificity of effects based on response habituation alone.

To localize potential convergence zones for processing of dynamic social information (i.e., emotional cues in facial expressions and prosody) irrespective of the sensory modality, we complemented the classical fMRI adaptation method with analysis strategies that enabled us to identify brain networks in which activity is correlated with behavioral effects (i.e., faster reaction times during classification of social cues) as well as structural approaches that reveal fiber connections towards these areas. Specifically, we hypothesized structural fiber projections between the areas for processing of faces, voices, and audiovisual integration along the right STS on the one hand and the OFC as a multisensory convergence zone for processing of dynamic facial and vocal cues on the other hand. Moreover, we predicted that modality-specific cortices, such as the face-sensitive and voice-sensitive cortex in the STS, would show an enhanced response habituation for their respective preferred stimulus class (i.e., stronger habituation of face-sensitive STS to faces than voices and vice versa for the voice-sensitive STS), whereas multisensory convergence zones, such as the audiovisual integration area in the STS and the OFC, would elicit a consistent habituation pattern that does not depend on the sensory modality occurring similarly for faces, voices, and face–voice combinations.

**Material and methods**

**Participants, stimulus material, and experimental design**

Twenty-three healthy, right-handed German native speakers (13 females; 23.0 ± 4.2 years) participated in one DTI and three fMRI experiments. Right-handedness was assessed with the Edinburgh Inventory (Oldfield, 1971). None of the participants had a history of neurological or psychiatric illness, substance abuse, or impaired hearing. Vision was normal or corrected to normal. None of the participants was on any medication. The study was performed according to the Code of Ethics of the World Medical Association (Declaration of Helsinki). All subjects gave their written informed consent prior to inclusion in the study.

The fMRI experiments included a face localizer (Kanwisher et al., 1997), a voice localizer (Belin et al., 2000) and a bimodal face–voice integration localizer (Kreifelts et al., 2009) modified to additionally enable investigation of modality-dependent habituation effects (Ethofer et al., 2009a; Grill-Spector et al., 1999).

The face localizer was adapted from previous studies on face processing (Epstein et al., 1999; Haxby et al., 2000; Kanwisher et al., 1997) and included pictures from four different categories (faces, houses, objects, and natural scenes) presented using a block-design. Six blocks (duration: 16 s) of each category pseudorandomized within the experiment were presented separated by short rest periods of 1.5 s. Within each block, 20 stimuli of one category were presented in a random order for 300 ms interleaved with 500 ms of fixation. To keep the participants’ attention fixed on the stimuli, they were instructed to press a button with their right index finger when they saw a picture directly repeated (one-back task). Positions of repeated stimuli were randomized within blocks with the restriction that one occurred during the first half of the block and one during the second half.

The voice localizer consisted of a passive-listening block design experiment with 32 stimulations and 16 silent epochs (each 8 s), as validated in previous research (Belin et al., 2000; Ethofer et al., 2009b). These stimuli included 16 blocks with human voices (HV; e.g., speech, sighs, laughs), eight blocks with animal sounds (AS; cries of various animals), and eight blocks with environmental sounds (ES; e.g., doors, telephones, cars).

The stimulus material used in the face–voice integration experiment consisted of short video clips (duration: 848 ms ± 295 ms, mean ± standard deviation) during which professional actors spoke words with emotionally neutral semantic content (the list of words is presented as supplemental material) in happy, neutral, or angry prosody with an emotionally congruent facial expression. These stimuli were evaluated in a prestudy outside the scanner including 20 subjects (10 females, 24.2 ± 3.8 years) to confirm that the emotional category intended by the actors was recognized with a high accuracy during auditory (A = sound clip without visual presentation; mean recognition rate: 84%) visual (V = mute video clip; mean recognition rate: 89%) and audiovisual presentation (AV = video clip with sound; mean recognition rate: 96%). To investigate modality-dependent habituation effects, each stimulus was presented three times during the course of the experiment (18 words × 3 modalities × 3 repetitions = 162 stimuli). These stimuli were presented within the framework of an event-related design with a varying inter stimulus interval (8.2–10.2 s) during three imaging runs (duration: about 10 min) each of which contained 18 auditory, 18 visual, and 18 audiovisual stimuli. On average, repetitions occurred with a temporal delay of 47.6 ± 36.9 s and 5.2 ± 4.0 intervening stimuli for all three modalities. To
maximize behavioral priming and repetition suppression effects, we limited the maximal time interval between the repetitions by clustering the stimuli presented in each of the three imaging runs in three segments consisting of 18 trials. Each of the words used as stimuli was presented in exactly one modality (A, V, or AV) during each of the three imaging runs. The order of the three runs was counterbalanced across subjects. Attentional processing of the stimuli during the fMRI experiment was ensured by instructing the participants to determine the expressed emotional category as quickly as possible by pressing one of three buttons in their right hand.

Image acquisition

Structural $T_1$-weighted images ($TR = 2300$ ms, $TE = 2.96$ ms, $TI = 1100$ ms, voxel size: $1 \times 1 \times 1$ mm$^3$) and functional images ($30$ axial slices acquired in sequential descending order, slice thickness $4$ mm $+ 1$ mm gap, $TR = 1.7$ s, $TE = 30$ ms, voxel size: $3 \times 3 \times 5$ mm$^3$) were acquired with a 3 T scanner (Siemens TIM TRIO, Erlangen, Germany). Time series consisted of 331 images for the face localizer, 226 images for the voice localizer, and $3 \times 330$ images for the bimodal face–voice localizer. For correction of image distortions, a field map (36 slices, slice thickness $3$ mm $+ 1$ mm gap, $TR = 400$ ms, $TE(1) = 5.19$ ms, $TE(2) = 7.65$ ms, voxel size: $3 \times 3 \times 4$ mm$^3$) was acquired. Diffusion-weighted images were acquired using a “Stejskal–Tanner” sequence ($TR = 8.3$ s, $TE = 82$ ms, flip angle = 90°, 64 axial slices, 2 acquisitions) with a voxel size of $2 \times 2 \times 2$ mm$^3$ along 30 independent directions using a $b$-value of 1000 s/mm$^2$. Additionally, an image with a $b$-value of 0 s/mm$^2$ was acquired for coregistration with the fMRI data.

Analysis of behavioral data

To determine the participants’ efficacy in classifying the stimuli depending on sensory modality and repetition, we measured reaction times and determined the unbiased hit rate (Hu) which was devised to include false alarms and biases in the use of response categories in the analysis of nonverbal behavior. This measure is defined as “the joint probability that a stimulus category is correctly identified given that it is presented at all and that a response is correctly used given that it is used at all” (Wagner, 1993). The Hu is obtained by multiplying the raw hit rate with the positive predictive value and does, thus, not only capture how sensitive, but also how specific a categorization task is carried out. Reaction times and Hu values were submitted to a two-way analysis of variance (ANOVA) for repeated measures with modality (A, V, AV) and presentation (first, second, third) as within-subject factors. All resulting $P$ values were corrected for heterogeneous correlations (Geisser and Greenhouse, 1958). Post-hoc $T$-tests were calculated to compare behavioral responses across modalities (A vs. V, A vs. AV, and V vs. AV) and $T$ values were corrected for multiple comparisons using a Bonferroni correction.

Analysis of fMRI data

Images were analyzed with statistical parametric mapping software (SPM5, Wellcome Department of Imaging Neuroscience, London, UK). Preprocessing comprised realignment, unwarping to correct for field distortions and to remove residual movement-related variance due to interactions between motion and field distortions (Andersson et al., 2001), slice time correction, normalization into MNI space (Montreal Neurological Institute, resampled voxel size: $3 \times 3 \times 3$ mm$^3$), and smoothing with a Gaussian filter (10 mm full width at half maximum).

For the face and voice localizer experiment, responses to the single categories (faces, houses, objects, and scenes in the face localizer and HV, AS, and ES in the voice localizer) were separately modeled using a box-car function corresponding to the block duration (16 s in the face localizer and 8 s in the voice localizer) convolved with the hemodynamic response function (hrf). For the face–voice integration experiment, responses to A, V, and AV stimuli were modeled separately for each trial as event-related responses using a stick function convolved with the hrf.

For all three experiments, data from the individual first-level general linear models were employed to create contrast images for each subject which were then submitted to a second-level random-effect analysis to enable population inference. The following contrasts were obtained using conjunction analyses based on the conjunction null hypothesis (Nichols et al., 2005): Experiment 1: Face sensitivity (Faces $>$ Houses $\cap$ Faces $>$ Objects $\cap$ Faces $>$ Scences), experiment 2: Voice sensitivity (HV $>$ AS $\cap$ HV $>$ ES), experiment 3: Audiospatial integration (AV $>$ A $\cap$ AV $>$ V). The selection of appropriate statistical criteria for defining multisensory integration is intensively discussed in the neuroimaging community (e.g., Beauchamp, 2005; Calvert, 2001; Goebel and van Anteveldt, 2009; Laurienti et al., 2005; Love et al., 2011; Stein et al., 2009). To date, no single criterion can be claimed to be the optimal solution. The criteria receiving most attention within this discussion are the maximum criterion (i.e., significantly stronger activation in the bimodal as compared to both unimodal conditions separately) and the superadditivity criterion (i.e., stronger activation in the bimodal condition as compared to the sum of activations in both unimodal conditions). In our study, the definition of audiovisual integration areas relied on the maximum criterion rather than the superadditivity criterion, since the latter can produce activation maps that are at least partly driven by deactivations in one or both of the unimodal conditions. Such mathematical superadditivity effects have been observed for the parietal regions (e.g., Brefczynski-Lewis et al., 2009; Ethofer et al., 2006c; Joassin et al., 2011a, 2011b) which converges with ERP findings (e.g. Brefczynski-Lewis et al., 2009; Joassin et al., 2004). Given the dependency of such effects on deactivations in the unimodal conditions and the fact that neuroimaging methods sample activity from large neural populations, however, the applicability of the superadditivity criterion which was originally developed on the basis of single cell recordings to fMRI or ERP data is an ongoing debate (Beauchamp, 2005; Ethofer et al., 2006c; Klemen and Chambers, 2012; Laurienti et al., 2005).

To investigate modality-dependent habituation effects, brain activations obtained for the first and third representation were separately compared for V, A, and AV stimuli and a conjunction analysis depicting habituation effects across modalities was carried out ($V_1 > V_3 \cap A_1 > A_3 \cap AV_1 > AV_3$). Furthermore, a simple regression analysis between single trial responses during V, A, and AV stimulation and the corresponding reaction times was performed to investigate in which brain areas the activation shows a linear relationship with behavioral facilitation due to repetition priming effects. Again, a conjunction analysis was employed to determine whether such relationships between brain responses and reaction times occur irrespective of the sensory modality. All fMRI activation maps are reported using a height threshold of $p < 0.001$ (uncorrected) and an extent threshold of $k > 15$ voxels. Correction for multiple comparisons ($p < 0.05$, corrected) across the whole brain was assessed at cluster level using random field theory ($k > 60$ voxels). Assignment of brain structures to activation clusters was done using the automatic anatomic labeling atlas (AAL, Tzourio-Mazoyer et al., 2002).

Average responses (mean $\pm$ standard error of beta values) obtained from the cortex showing face-sensitivity, voice-sensitivity, and audiovisual face–voice integration along the right STS as well as brain areas showing significant habituation effects across sensory modalities (activation clusters obtained at group level) were submitted to a two-factorial repeated measures ANOVA with modality (A, V, AV) and presentation (first, second, third) as within-subjects factor. As the cluster for face–voice integration in the STS overlapped with the clusters showing face sensitivity and voice sensitivity (see Results), the activation data of the latter two included only voxels which did not show a significant audiovisual integration effect. We additionally...
employed paired T tests to investigate whether the STS regions habituate most strongly to their preferred stimulus category. To this end, the habituation (first presentation > third presentation) was statistically compared across modalities.

Analysis of DTI data

Diffusion-weighted data were analyzed using FSL 4.04 (FMRIB Software Library, Oxford University, www.fmrib.ox.ac.uk/fsl). Preprocessing of diffusion-weighted images included eddy current correction and averaging across the two acquisitions. Voxel-wise estimates of fiber orientations and their uncertainty were calculated on the basis of a model that accounts for the possibility of crossing fibers within each voxel (Behrens et al., 2003, 2007). The two most probable directions within each voxel of the white-matter were determined for each individual subject.

Three different seed masks along the right STS were generated using the activation clusters obtained from the fMRI experiments for face sensitivity (Faces > Houses ∩ Faces > Objects ∩ Faces > Scenes), voice sensitivity (HV > AS ∩ HV > ES), and auditory–visual integration (AV > A ∩ AV > V). As the cluster for bimodal face–voice integration overlapped with face sensitivity and voice sensitivity (see Results), the seed masks of the latter two only included voxels which did not show a significant auditory–visual integration effect. In addition, we investigated the connectivity profile of brain areas showing significant habituation effects across modalities (V1 > V3 ∩ A1 > A3 ∩ AV1 > AV3) to verify structural fiber connections with the STS regions for face and voice processing as well as auditory–visual face–voice integration. All employed seed masks contained only voxels within the grey matter (determined by segmentation of the high resolution anatomical images, Ashburner and Friston, 2005) which overlapped with activation clusters obtained in the fMRI group analyses. These seed masks were transferred to individual DTI space (resampled voxel size: 2 × 2 × 2 mm³). Probabilistic fiber tracking was not restricted by target or waypoint masks. Probabilistic fiber tracking maps were thresholded at 5% of the maximum value to reduce false-positive fiber tracks. These thresholded probabilistic fiber tracking maps were then back-transformed to MNI space, and binary connectivity maps (containing ones in voxels where connections were found and zeros elsewhere) were generated for each subject. These individual binary connectivity maps were added across subjects and fiber projections present in more than 50% of the subjects were displayed on the border between grey and white matter of the 3D mean normalized T1-weighted image of the participants using the freely available POV-Ray software (www.povray.com). In addition, we performed a conjunction analysis (Nichols et al., 2005) to depict common fiber projections of the face- and voice-sensitive regions as well as the audiovisual face–voice integration area within the cortex of the right STS. To locate the predominant pathways for each region, we corrected the probabilistic fiber tracking results for its size (i.e., divided by the number of seed voxels) and statistically compared the connectivity maps for each STS area with the respective other two regions (e.g., fiber tracts of face-sensitive STS area > fiber tracts of audiovisual face–voice integration STS area and fiber tracts of face-sensitive STS area > fiber tracts of voice-sensitive STS area) using paired t-tests and a conjunction analysis based on the minimum T statistic (Nichols et al., 2005). In analogy to the fMRI analysis, correction for multiple comparisons relied on random-field theory (Worsley et al., 1996) using an extent threshold of k > 60 voxels corresponding to p < 0.05, corrected for multiple comparisons across the whole brain.

Results

Behavioral data

Similar raw hit rates as in the behavioral prestudy were found for classification of auditory (78.1 ± 1.6%), visual (88.8 ± 1.1%), and audiovisual (93.9 ± 3.6%) trials indicating attentive processing of the stimuli throughout the fMRI experiment. The two-factorial repeated measures ANOVA on unbiased hit rates revealed a main effect of modality (F(2.44) = 59.5; p < 0.001). This main effect was due to higher recognition rates for visual than for auditory stimuli (paired T(22) = 6.1; p < 0.001) as well as higher recognition rates for audiovisual stimuli as compared to both auditory (paired T(22) = 10.9; p < 0.001) and visual (paired T(22) = 4.8; p < 0.001) stimuli. The main effect of repetition was also significant (F(2.44) = 23.9; p < 0.001) which was driven by significantly higher recognition rates in the second and third as compared to the first presentation in all three modalities (all paired T(22) > 2.4; all p < 0.05). No significant differences in recognition rates were found when comparing the second and third presentation. Furthermore, no interaction between modality and repetition (F(4.88) = 0.9; p > 0.05) was found as repetition improved recognition rates to a similar extent for auditory, visual, and audiovisual stimuli (see Fig. 1a). Analysis of reaction times revealed a strong main effect of repetition (F(2.44) = 2237.4, p < 0.001) with acceleration of behavioral responses between the first and second presentation (all paired T(22) > 21.5, all p < 0.001) as well as second and third presentation (all paired T(22) > 10.0, all p < 0.001) indicating reliable repetition priming across modalities (see Fig. 1b). Furthermore, the main effect of modality (F(2.44) = 1198.6; p < 0.001) was also significant with shorter reaction times during perception of audiovisual stimuli as compared to both unimodal visual and auditory stimuli (paired T(22) > 42.9; p < 0.001) as well as faster responses for visual than auditory stimuli (paired T(22) = 4.6, p < 0.05). No interaction between modality and repetition was found (F(4.88) = 0.5; p > 0.05).

fmRI data

Face-sensitive regions as defined by the conjunction analysis (Faces > Houses ∩ Faces > Objects ∩ Faces > Scenes, see Fig. 2a) were situated in the right pSTS and bilateral precuneus/posterior
cingulate cortex (PC/pCC). In addition, this contrast yielded activations in the left pSTS, right fusiform gyrus (FG), and right amygdala (AMY). However, these activations failed to reach the significance after correction for multiple comparisons across the whole brain and are thus reported descriptively here. The conjunction analysis (HV > AS ∩ HV > ES, see Fig. 2b) revealed voice-sensitive areas in bilateral mSTS. Brain areas showing audiovisual integration effects (AV > A ∩ AV > V, see Fig. 2c) included the right pSTS at the overlap of the face- and voice-sensitive cortex as well as the bilateral posterior thalamus. Of these areas, only the right pSTS was significant after correction for multiple comparisons across the whole brain. In the left pSTS, audiovisual integration effects (MNI coordinates x = −57; y = −51; z = 0; Z score = 2.83) failed to reach the predefined height threshold of p < 0.001. Activation maxima and cluster sizes of brain areas showing face-sensitivity, voice-sensitivity or audiovisual face–voice integration are given in Table 1. Brain responses to the first, second and third presentation during perception of A, V, and AV stimuli are presented in Fig. 2 for the cortex along the right STS separately for the face-sensitive area (as defined by the contrast: Faces > Houses ∩ Faces > Objects ∩ Faces > Scenes, Fig. 2a), voice-sensitive area (as defined by the contrast: HV > AS ∩ HV > ES, Fig. 2b), and the audiovisual face–voice integration area (as defined by the contrast AV > A ∩ AV > V, Fig. 2c).

The two-factorial repeated measures ANOVA on responses of the face-sensitive cortex in the right STS yielded a main effect of modality (F(2,44) = 11.8, p < 0.01) and a main effect of repetition (F(2,44) =

Fig. 2. Areas along the right STS (p < 0.001, uncorrected, k > 15 voxels) showing face sensitivity (a), voice sensitivity (b), and audiovisual integration effects (c) as displayed on sagittal slices as well as the corresponding activation strengths (mean ± standard error) during the first, second, and third presentation of visual (white bars), auditory (grey bars), and audiovisual stimuli (black bars).
8.7, p < 0.01). The interaction modality × repetition failed to reach significance (F(4,48) = 2.0; p = 0.10). Direct comparison of habituation effects in the face-sensitive area revealed a stronger habituation during repeated unimodal visual than unimodal auditory stimulation (paired T(22) = 1.9, p < 0.05). For the voice-sensitive cortex within the right mSTS, a significant main effect of modality (F(2,44) = 201.6, p < 0.001), a main effect of repetition (F(2,44) = 22.1, p < 0.001), and a significant interaction modality × repetition (F(4,48) = 3.1; p < 0.05) was found. Statistical comparison of habituation indicated significantly stronger repetition suppression during unimodal acoustic than unimodal visual presentation (paired T(22) = 1.9, p < 0.05). For the face–voice integration area, a main effect of repetition (F(2,44) = 14.0, p < 0.001), but no interaction (F(4,48) = 1.2; p = 0.31) was found. Habituation to AV stimuli in this area was not significantly different to that obtained for unimodal stimulation.

Significant habituation effects occurring for all three conditions (conjunction analysis: V1 > V3 & A1 > A3 & AV1 > AV3, see Fig. 3a, Table 2) were found bilaterally within the orbital part of the inferior frontal gyrus (orbIFG). The two-factorial ANOVA on responses within these areas revealed a main effect of repetition in both hemispheres (right orbIFG: F(2,44) = 14.0, p < 0.001, left orbIFG: F(2,44) = 38.8, p < 0.001). For the left orbIFG, a main effect of modality (F(2,44) = 10.3, p < 0.01) with stronger responses to unimodal auditory than both unimodal visual and bimodal stimulation was observed. No main effect of modality was found for the right orbIFG (F(2,44) = 2.9, p = 0.06). Most importantly, however, habituation of brain responses was similar for A, V, and AV in bilateral orbIFG (see Fig. 3b, c) and consequently there was no interaction between modality and repetition (right orbIFG: F(4,48) = 1.7; p = 0.15, left orbIFG: F(4,48) = 0.4; p = 0.77).

The results of the simple regression analysis between brain activity and reaction times are presented in Fig. 4 and Table 3. A significant linear relationship for V, A, and AV stimuli was found in the bilateral orbIFG (right orbIFG: V: r = 0.29 ± 0.04, A: r = 0.31 ± 0.03, AV: r = 0.25 ± 0.03; left orbIFG: V: r = 0.28 ± 0.04, A: 0.30 ± 0.04; AV: 0.21 ± 0.03), as well as the motor system including the supplementary motor area (SMA, V: r = 0.29 ± 0.05, A: r = 0.30 ± 0.04, AV: r = 0.21 ± 0.04, all values in mean ± standard error) and bilateral caudate nucleus.

### Discussion

In the current study, we employed an MRI adaptation design that allowed us to investigate neural networks subserving sensory-related processing (i.e., areas with adaptation of their responses during repeated stimulus exposure) as well as brain areas involved in generation of faster behavioral responses (i.e., areas where activity is correlated with reaction times on a trial by trial basis). Furthermore, we combined this method with probabilistic fiber tracking to elucidate the structural connectivity architecture of these systems. Our findings enable novel insights into the brain anatomy underlying social signals conveyed by facial expressions and prosody and provide constraints for models of neural pathways underlying face and voice processing.

### Habitation of brain responses to social signals in faces and voices

The anatomical locations of activation clusters along the right STS were in agreement with previous neuroimaging studies on processing of faces (e.g., Fox et al., 2009a; Pourtois et al., 2009), voices (e.g., Belin et al., 2000; Ethofer et al., 2009a), and face–voice combinations (e.g., Kreifelts et al., 2009). A detailed analysis of the differential contributions of anatomical structures to the activated regions (see Table 1)
revealed that the face-sensitive area was mostly situated within the middle temporal gyrus (MTG) while the largest part of the voice-sensitive area was located in the superior temporal gyrus (STG) which is in line with the parcellation of the STS cortex in macaque monkeys (Seltzer and Pandya, 1978) demonstrating that visual and auditory areas are predominantly located in the lower and upper banks of the STS, respectively. We used standard face and voice localizers to maximize comparability and generalizability of our findings. The employed face localizer relied on static faces. Thus, brain areas that show a strong preference for dynamic face stimuli, but elicit only a weak response to static faces (e.g., in the anterior part of the STS, Pitcher et al., 2011) were not included by this approach. Furthermore, it should be noted that the standard voice localizer (Belin et al., 2000) also contains some emotional stimuli in the voice condition (e.g., laughs and cries). However, the same is true for the animal voice condition (e.g., trumpeting elephant or cat purring of pleasure in the animal voice condition) and the environmental sounds (e.g., a gunshot or a police siren). Therefore, it is unlikely that the results are biased by different amounts of emotion. The same is true for the face–voice integration localizer which was based on the same number of emotional and neutral stimuli in all three conditions.

We presented our stimuli within the framework of a factorial habituation design to investigate repetition priming effects during processing of social stimuli across sensory modalities. Behavioral data demonstrated reliable facilitation of responses during repeated stimulus exposure. In line with these observations, responses in face-sensitive areas in the right pSTS habituated stronger to faces than voices while voice-sensitive areas in the right mSTS exhibited the opposite effect indicating enhanced sharpening of neural responses within these areas to stimuli of their preferred categories. Within the audiovisual integration area, repetition suppression effects were similar for faces, voices, and face–voice combinations. These findings concur with neurophysiological observations in macaques indicating that only a minority of cells (about 20%) in the superior temporal polysensory area (STP) is sensitive to both auditory and visual information (Bruce et al., 1981) and shows a modulation of their responses to facial gestures if they are accompanied by a matching sound (Barraclough et al., 2005). Moreover, neuroimaging data using high resolution fMRI demonstrated a patchy organization of these cells within the pSTS (Beauchamp et al., 2004). Similar results have been obtained from audiovisual integration of letters and sounds demonstrating a distributed pattern of STS areas that preferentially habituate to congruent or incongruent stimuli (van Atteveldt et al., 2010). Future studies combining such high resolution techniques with an adaptation design should address the question whether clusters showing maximal responses to audiovisual stimuli exhibit a specific repetition suppression during perception of face–voice combinations indicating a specific tuning of these cortical patches to multimodal stimulation.

We performed a conjunction analysis to identify which other brain areas habituate their response during repeated stimulus exposure irrespective of the sensory modality. The only brain regions showing significant response attenuation across conditions were found in the orbIFG of both hemispheres indicating a tuning of neural responses to facial and vocal cues. This bilateral activation pattern at the level of the frontal cortex during processing such signals accords with previous neuroimaging data on comprehension of social signals in prosody (e.g., Ethofer et al., 2006a; Wildgruber et al., 2004) and facial expressions (for a recent meta-analysis, see Sabatinelli et al., 2011). Furthermore, these data are also in line with imaging data obtained using combined EEG–fMRI demonstrating consistent activation within bilateral inferior frontal cortex during perception of voices, faces, and face–voice combinations (Brefczynski-Lewis et al., 2009). Finally, our data converge with neuropsychological observations demonstrating deficits in recognition of social signals from voice and face (Hornak et al., 1996, 2003) as well as reduced social functioning (Beer et al., 2006) in patients with lesions in either left or right orbitofrontal cortex (OFC).

### Table 2
Brain areas showing habituation effects across sensory modalities.

<table>
<thead>
<tr>
<th>Brain area</th>
<th>MNI coordinates</th>
<th>Z score</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right inferior frontal gyrus, orbital part</td>
<td>30 33 − 12</td>
<td>4.12</td>
<td>192*</td>
</tr>
<tr>
<td>Left inferior frontal gyrus, orbital part</td>
<td>−39 24 0</td>
<td>4.27</td>
<td>165*</td>
</tr>
<tr>
<td>Right inferior frontal gyrus</td>
<td>45 12 27</td>
<td>3.95</td>
<td>57</td>
</tr>
<tr>
<td>Left fusiform gyrus</td>
<td>−42 − 51 − 18</td>
<td>3.70</td>
<td>29</td>
</tr>
<tr>
<td>Left inferior frontal gyrus</td>
<td>−39 6 27</td>
<td>3.50</td>
<td>24</td>
</tr>
<tr>
<td>Left caudate nucleus</td>
<td>−9 8 0</td>
<td>3.75</td>
<td>22</td>
</tr>
</tbody>
</table>

* p < 0.05, corrected at cluster level (k > 60 voxels) across the whole brain.
In analogy to the habituation profile found for the audiovisual face–voice integration area in the STS, response habituation in the orbIFG was similar for perception of voices, faces, and face–voice combinations. An intertwined and overlapping distribution of face-responsive, voice-responsive and multisensory neurons, as described for this area in macaque monkeys (Romanski, 2012), offers an explanation for this similarity of habituation across sensory modalities.

The strength of the current study is that the conjunction approach enabled us to provide evidence for adaptation of responses in bilateral orbIFG during repeated exposure of dynamic signals in the context of emotion categorization task which occurred irrespective of the sensory modality. It is, however, possible that the orbIFG is also involved

### Table 3

<table>
<thead>
<tr>
<th>Brain area</th>
<th>MNI coordinates</th>
<th>Z score</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left inferior frontal gyrus, orbital part</td>
<td>−39 24 6</td>
<td>4.40</td>
<td>356*</td>
</tr>
<tr>
<td>Bilateral supplementary motor area</td>
<td>−6 24 45</td>
<td>4.40</td>
<td>352*</td>
</tr>
<tr>
<td>Right inferior frontal gyrus, orbital part</td>
<td>33 27 0</td>
<td>4.03</td>
<td>140*</td>
</tr>
<tr>
<td>Right caudate nucleus</td>
<td>12 9 0</td>
<td>3.83</td>
<td>91*</td>
</tr>
<tr>
<td>Right inferior frontal gyrus</td>
<td>54 24 24</td>
<td>3.76</td>
<td>51</td>
</tr>
<tr>
<td>Left inferior frontal gyrus</td>
<td>−42 0 39</td>
<td>3.75</td>
<td>41</td>
</tr>
</tbody>
</table>

* p < 0.05, corrected at cluster level (k > 60 voxels) across the whole brain.

---

Fig. 4. Simple regression analysis of brain activation and reaction times (p < 0.001, uncorrected, k > 15 voxels) rendered on sagittal and coronal slices (a). Structural connections of the right orbIFG consistently found in more than half of the subjects are shown in blue/light blue. Data extracted from one typical subject for the left orbIFG (left panel), supplemental motor area (middle panel), and right orbIFG (right) separately for visual (b), auditory (c), and audiovisual trials (d).
in processing of other classes of stimuli (e.g., audiovisual speech, Hasson et al., 2007) and, thus, future research is needed to compare repetition suppression effects to faces and voices with non-social auditory and visual stimuli or under different task instructions (e.g., gender or identity judgment, Joassin et al., 2011a; Joassin et al., 2011b).

Correlation of brain responses with behavioral facilitation during repetition priming

To directly study the neural systems subserving the behavioral output during perception of social information in voices and faces, we conducted a simple regression analysis using the participants’ reaction times during judgment of the perceived emotional category. A significant correlation was found for the bilateral orbIFG indicating a pivotal role of this area for mediating facilitation of responses during repetition priming. In addition, components of the motor system including the SMA and caudate nucleus also showed a linear relationship between their response amplitudes and reaction times. These findings nicely converge with previous fMRI data demonstrating that activity in these areas predicts behavioral responses during auditory signal detection (Sadaghiani et al., 2009) and suggest that the orbIFG might serve as a computational hub within the intrinsic alertness

Fig. 5. Structural connection profile of the face-sensitive (a), voice-sensitive (b), and audiovisual integration areas (c) along the right STS and conjunction of projections from these three brain areas (d) displayed on the border between grey and white matter of the 3D mean normalized T1-weighted image. Brain regions showing consistent fiber connections in more than half of the participants are shown in blue (left panels). Statistical comparison of fiber pathways revealed predominant projections along the dorsal SLF for the face-sensitive (red), EC for voice-sensitive (green), and ventral SLF audiovisual integration areas (yellow).
system (Fox et al., 2005, 2009b; Fransson, 2005) linking sensory cortices and motor systems during active judgment of socially relevant information.

**Structural connectivity of brain areas for processing and integration of faces and voices**

To directly investigate such structural connections, we conducted a probabilistic fiber tracking analysis using the activation clusters along the right STS that exhibited a significant sensitivity to faces, voices or face–voice combinations. In agreement with our predictions, fibers originating in the face sensitive area in the right pSTS projected to the orbIFG as well as the adjacent insula. Statistical comparison of fiber bundles that started in the voice-sensitive cortex or the neighboring audiovisual integration area revealed that structural connections towards the IFG were most prominently running through the dorsal part of the SLF. These direct connections nicely dovetail with electrophysiological findings obtained in macaques suggesting subsequent processing steps in the time range of 50–170 ms in the STS (e.g., Keyser et al., 2001) and 100–340 ms in the OFC (e.g., Rolls et al., 2006) during perception of faces (for a review, see Barracough and Perrett, 2011). No direct fiber connections between the right pSTS and FG were found confirming previous fiber tracking results (Gschwind et al., 2012). Instead, the face sensitive area showed consistent fiber projections via the ILF that terminated in the medial part of the inferior temporal cortex near the temporal pole. This area is known to be well-connected with the extrastriate visual cortex (Catani et al., 2003) and neuropsychological data from lesion studies (Damasio et al., 1990) as well as neuroimaging data (Nakamura et al., 2000; Rotstein et al., 2005) suggest a strong commitment of this area in facial person recognition. Thus, the direct fiber connection between the pSTS and the anterior temporal lobe might constitute a neural substrate for integration of changeable and invariant facial features, such as emotional expressions and identity. The voice area was strongly connected with the orbIFG and the parietal lobe confirming previous tracking results obtained for the connectivity profile of voice-sensitive areas that respond to emotional prosody (Ethofer et al., 2012). These findings are in line with current models on auditory perception suggesting separate processing streams for spatial and frequency-related information (i.e., ‘where’ versus ‘how’ pathways, Belin and Zatorre, 2000). In agreement with previous DTI results (Blank et al., 2011), short range connections of the voice area reached the anterior part of the lateral temporal lobe, an area which is thought to be involved in vocal speaker recognition (von Kriegstein et al., 2003) and directly interacts with the FFA during bimodal speaker recognition (von Kriegstein and Giraud, 2006). A recently published study described direct fiber connections between voice areas in the STG and face sensitive areas in the FG (Blank et al., 2011). No such connections were found in the present study. However, it should be noted that the voice areas in the study of Blank and colleagues were defined by a task-related contrast (comparison of speaker and speech identification) and the strongest connections in that study were found for the anterior temporal lobe. This area was not included in our seed region which was obtained by a stimulus-related contrast (human voices versus animal voices and environmental sounds) providing a possible explanation for this discrepancy in results. The audiovisual integration area in right pSTS was found to be reliably connected to the orbIFG and statistical comparison of its projections with those of the neighboring face-sensitive and voice-sensitive areas revealed that this pathway was predominantly running through the ventral part of the SLF. Connections between the multimodal pSTS and the orbIFG are well in line with findings from anatomical studies in non-human primates providing evidence for input from several sensory modalities (Jones and Powell, 1970) and a pivotal role of orbitofrontal cortices for evaluation of social information (Beer et al., 2006). In a previous study (Ethofer et al., 2011), we demonstrated that STS areas sensitive to eye motion are also structurally connected with the orbIFG in the majority of subjects. However, in this study the most consistent connections were found for the anterior insula. These findings suggest that differences in connectivity profiles of neighboring or even partially overlapping brain regions (such as the area for audiovisual face–voice integration or eye gaze which are both located close to the bifurcation of the STS) can be determined. However, this has to be experimentally confirmed by studying the structural fiber connections of these areas in the same group of participants. The fiber tracking analysis using the right orbIFG as seed area confirmed consistent projections to the cortex along the STS. Furthermore, the right orbIFG additionally showed strong structural connections with critical components of motor circuits, such as the caudate nucleus and the SMA. Thus, our findings suggest that this area might constitute a critical neural hub for connecting systems required to generate behavioral responses to social information across sensory modalities.

**Conclusions**

Our study provides a first delineation of both the structural and functional substrates underlying perception of social signals in facial expressions and prosody. Using established fMRI localizers, we successfully defined areas for processing of faces and voices as well as their audiovisual integration along the right STS. Face and voice regions exhibited a stronger response habituation to their favored stimulus classes. No such differences in habituation was found for the audiovisual integration area which is in line with neurophysiological data indicating that only a minority of cells in this region is specifically tuned to bimodal stimuli. Probabilistic fiber tracking yielded converging structural connectivity of the three STS regions to the orbIFG which additionally showed consistent response habituation irrespective of the sensory modality. These combined structural and functional data indicate a key role for the orbIFG for processing of social signals conveyed by prosody and facial expressions and propose that this area is part of the extended system for both face and voice perception. Finally, regional activity in both orbIFG and SMA showed a linear relationship with reaction times during active judgment of social signals and probabilistic fiber tracking revealed strong structural connections between these areas. Based on these functional and structural properties, we propose that the orbIFG constitutes a convergence zone for face and voice processing linking sensory areas along the STS with motor areas that generate behavioral responses during judgment of social signals in voice and face.

**Funding**

This work was supported by grants of the Fortüne-Program of the University of Tübingen (fortüne 1874-0-0, fortüne 1997-0-0, and fortüne 2000-0-0).

**Conflict of interest statement**

The authors declare that there is no conflict of interest.

**Appendix A. Supplementary data**

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.02.064.

**References**


