How the Human Brain Exchanges Information Across Sensory Modalities to Recognize Other People

Helen Blank,^{1,2*} Stefan J. Kiebel,^{3,4,5} and Katharina von Kriegstein^{1,6}

¹Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

²MRC Cognition and Brain Sciences Unit, Cambridge CB2 7EF, United Kingdom

³Department of Neurology, Biomagnetic Centre, University Clinic Jena, 07747 Jena, Germany

⁴Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

⁵Department of Psychology, Technische Universität Dresden, 01069 Dresden, Germany

⁶Department of Psychology, Humboldt University of Berlin, 12489 Berlin, Germany

◆=====◆

Abstract: Recognizing the identity of other individuals across different sensory modalities is critical for successful social interaction. In the human brain, face- and voice-sensitive areas are separate, but structurally connected. What kind of information is exchanged between these specialized areas during cross-modal recognition of other individuals is currently unclear. For faces, specific areas are sensitive to identity and to physical properties. It is an open question whether voices activate representations of face identity or physical facial properties in these areas. To address this question, we used functional magnetic resonance imaging in humans and a voice-face priming design. In this design, familiar voices were followed by morphed faces that matched or mismatched with respect to identity or physical properties. The results showed that responses in face-sensitive regions were modulated when face identity or physical properties did not match to the preceding voice. The strength of this mismatch signal depended on the level of certainty the participant had about the voice identity. This suggests that both identity and physical property information was provided by the voice to face areas. The activity and connectivity profiles differed between face-sensitive areas: (i) the occipital face area seemed to receive information about both physical properties and identity, (ii) the fusiform face area seemed to receive identity, and (iii) the anterior temporal lobe seemed to receive predominantly identity information from the voice. We interpret these results within a prediction coding scheme in which both identity and physical property information is used across sensory modalities to recognize individuals. *Hum Brain Mapp* 36:324–339, 2015. © 2014 Wiley Periodicals, Inc.

Key words: cross-modal priming; face recognition; person identity; voice recognition; multisensory

◆=====◆

*Correspondence to: Helen Blank; MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, UK.
E-mail: helen.blank@mrc.cbu-cam.ac.uk
Conflict of Interest: The authors declare no conflict of interest.

Received for publication 7 July 2014; Accepted 29 August 2014.

DOI: 10.1002/hbm.22631

Published online 13 September 2014 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Recognizing the identity of other individuals is critical for successful social interaction of many species [Adachi et al., 2007; Kondo et al., 2012; Proops et al., 2009; Sidtis and Kreiman, 2012; Sliwa et al., 2011]. One key component of this ability is to associate identity across different modalities such as associating a face with the voice, for example, if someone calls out “Hello” from behind [Kamachi et al., 2003; Mavica and Barenholtz, 2013]. This skill seems highly relevant for social interaction and is also present in other species: Monkeys, horses, and crows, preferentially look at the correct face identity when they hear the voice of a familiar conspecific [Beauchemin et al., 2011; DeCasper and Fifer, 1980; Kondo et al., 2012; Proops et al., 2009; Sai, 2005; Sliwa et al., 2011]. How is this so-called cross-modal individual recognition implemented in the brain?

In both the human and macaque brain, there are several specialized areas processing faces and voices [Freiwald et al., 2009; Perrodin et al., 2011; Rajimehr et al., 2009]. In humans, there are three distinct face-sensitive areas involved in face recognition: an area in the inferior occipital gyrus (occipital face area, OFA), in the fusiform gyrus (fusiform face area, FFA), and in the anterior temporal lobe (aTL). These areas play different roles in the hierarchy from processing of lower-level physical facial properties to processing of face identity [Bruce and Young, 1986; Haxby et al., 2000]. The OFA processes mostly facial properties [Pitcher et al., 2011; Rossion, 2008; Rotshtein et al., 2007a], the FFA processes both facial properties and face identity [Grill-Spector et al., 2004; Kriegeskorte et al., 2007; Nestor et al., 2011; Pourtois et al., 2005; Rotshtein et al., 2005; Xu et al., 2009], and the aTL processes face identity [Garrido et al., 2009; Kriegeskorte et al., 2007; Rajimehr et al., 2009; Rotshtein et al., 2005; Tanji et al., 2012]. For voice-identity processing, voice-sensitive areas have been identified in the middle/anterior superior temporal sulcus (STS) [Andics et al., 2010; Belin and Zatorre, 2003; Belin et al., 2011; von Kriegstein and Giraud, 2004].

Several studies have shown that voice- and face-sensitive areas are directly and indirectly connected, suggesting an information transfer between these regions [Blank et al., 2011; Ethofer et al., 2012; Focker et al., 2011; von Kriegstein et al., 2005, 2008]. However, is information actually exchanged between these regions during cross-modal recognition, and if so, what kind of information? In the present study, we used functional magnetic resonance imaging (fMRI) in humans to address this question by testing which visual areas are sensitive to certain kinds of

voice information during cross-modal recognition: (a) conceptual information about face identity or (b) physical properties of faces.

We derived our hypotheses about the neural mechanisms of cross-modal information transfer between voice- and face-sensitive areas from the predictive coding framework [Friston, 2005; Mumford, 1992; Rao and Ballard, 1999]. The predictive coding framework suggests that incoming information is processed in a hierarchical way so that information at lower processing levels is compared against predictions from higher levels; only the unpredicted information (i.e., the prediction error) is forwarded to higher levels. Therefore, unpredicted, mismatching information leads to an increase in measured activity and vice versa, high predictability leads to reduced activity. Here, we translate this view to the present cross-modal design; if information from voice-sensitive areas is used to predict the corresponding face, measured activity in face-sensitive regions should be higher for a mismatch, than for a match of voice and face. Given the connections between FFA and the voice-sensitive regions [Blank et al., 2011], we expected that during voice recognition, voice-sensitive areas in middle/anterior STS send predictions about identity and/or physical properties of the corresponding face to FFA. As the face-sensitive regions are interconnected regions [Gschwind et al., 2012; Mechelli et al., 2004; Thomas et al., 2009], this information is potentially further communicated within the face-sensitive network from FFA to OFA and possibly aTL. If this is the case, then the face-sensitive areas should produce prediction errors during the subsequent presentation of mismatching faces and we tested this in the present experiment. Using fMRI, we measured the difference in prediction errors as activity changes during the presentation of the mismatching face following the voice prime, and compared these to the presentation of the matching face. In addition, within the predictive coding framework, the amount of prediction error depends on the certainty of the prediction [Friston, 2005; Mumford, 1992; Rao and Ballard, 1999]. Accordingly, we expected that the strength of the measured prediction errors depended on the level of certainty the participant had about the identity of the voice, that is, how fast the voices were recognized.

The obtained results are congruent with a predictive coding scheme [Friston, 2005; Mumford, 1992; Rao and Ballard, 1999] where both informations about identity and physical properties could be used across sensory modalities to accomplish fast and robust recognition of individuals.

MATERIALS AND METHODS

We acquired fMRI data while participants performed identity recognition in a cross-modal priming experiment in which familiar voices preceded morphed faces (Fig. 1A). The morphed faces matched or mismatched the

Abbreviations

aTL	anterior temporal lobe
FFA	fusiform face area
fMRI	functional magnetic resonance imaging
OFA	occipital face area

Figure 1.

Experimental design and face stimuli. **A.** Participants were first trained in identifying the voices and faces of three male speakers (Training) and participated in a psychophysical pilot experiment (Pilot) (see Methods). Then, participants performed a cross-modal priming experiment and a face-area localizer during fMRI. In the cross-modal priming experiment, voices of three speakers (indicated by amplitude waveforms) were followed by images of their faces after a 75 ms delay. The voices and faces could match (light gray font) or mismatch (dark gray font). Because the faces were morphed continua between two face identities (see panel B), voice and face could match or mismatch with regard to identity or physical properties. Participants performed a face-identity recognition task on every trial (face task); on some trials they additionally performed a voice-identity recognition task (“voice task”; excluded from fMRI analysis; see Methods). This was indicated by a colored fixation cross that followed the voice presentation. The faces were morphed between the three speakers (see Methods, panel B). **B.** Morphed face stimuli of the three speakers: Each combination of two possible face-identity pairs was morphed, resulting in three morphed continua (Bernd–Anton, Anton–Carsten, and Carsten–Bernd). In the cross-modal priming experiment, the morph levels 0, 33, 67, and 100% were used. These levels parametrically differed in physical properties (light blue), but perception of identity differed in a categorical manner (dark blue, morph levels 0 and 30% were perceived as matching in identity to the voice and morph levels 67 and 100% were perceived as the other person, mismatching in identity) (see Methods).

preceding voice in relation to physical properties and/or identity [Kikutani et al., 2010; Rotshtein et al., 2005] (Fig. 1B). Within different face-sensitive areas, we tested whether functional activity and connectivity increased due to a mismatch (i) in face identity or (ii) in physical properties of faces. In the following, we describe the methods in detail.

Participants

Sixteen healthy volunteers (eight females; mean age 25 years, range of age 20–31 years; only right handers, assessed with the Edinburgh questionnaire [Oldfield, 1971]) participated in the study. Written informed consent was obtained from all participants according to procedures approved by the Research Ethics Committee of the University of Leipzig. One additional subject was excluded from the analysis because his performance in the voice recognition task was three standard deviations below the group average performance.

Stimuli

Face and voice recordings

Three male speakers’ faces and voices (22, 23, and 25 years old) were recorded with a digital video camera (Canon, Legria HF S10 HD-Camcorder). High quality auditory stimuli were simultaneously recorded with a condenser microphone (Neumann TLM 50, preamplifier LAKE PEOPLE Mic-Amp F-35, soundcard PowerMac G5, 44.1 kHz sampling rate, and 16-bit resolution) and the software Sound Studio 3 (Felt Tip). All recordings were done in a soundproof room under constant luminance conditions. All videos were processed and cut in Final Cut Pro (version 6, HD, Apple), converted to mpeg format and were presented at a size of 727 × 545 pixels. The high-quality audio-stream of each file was postprocessed using MATLAB (version 7.7, The MathWorks, MA) to adjust overall sound level to a root mean square of 0.083. Stimuli were presented and responses were recorded using Presentation software 14.1 (<http://www.neurobs.com/>).

Audio-visual samples (stimuli for the training)

To familiarize all participants with the three speakers, they were trained on audio-visual samples of the speakers’ faces and voices (Fig. 1A “training”; see below for details on the procedure). These audio-visual samples included semantically neutral and syntactically homogeneous five-word sentences (e.g., “Der Junge trägt einen Koffer.”; English: “The boy carries a suitcase.”). During video presentation the written name of each speaker (Anton, Bernd, and Carsten) was presented. To evaluate training success, single word voice samples and muted video samples of each speaker were used (e.g., “Dichter”; English: “poet”) (Fig. 1A).

Audio samples and morphed face images (stimuli for psychophysical pilot and fMRI experiments)

After the training, all participants performed first a psychophysical pilot experiment and then the cross-modal priming experiment in the MRI-scanner (Fig. 1A; see below for details on the procedure). Both these experiments included high-quality voice samples and face images taken as still frames from the video recordings (Fig. 1A,B).

The high-quality voice samples of each speaker included two-word sentences (the pronoun “er”/“he” and a verb; example: “Er kaut.”/“He chews.”).

Face images were morphed using FantaMorph (version 4.2.6, Abrosoft, <http://www.fantamorph.com>). Each combination of two possible face-identity pairs was morphed, resulting in three morphed continua (Anton to Bernd, Anton to Carsten, Bernd to Carsten). We produced morph levels from 0 to 100% in increments of 10% and additionally levels 33 and 67%. In the psychophysical pilot experiment, all morph levels were used, while the fMRI cross-modal priming experiment included levels 0, 33, 67, and 100% (Fig. 1B). Visual stimuli were presented at a visual angle of $10 \times 8^\circ$.

Face and object images (stimuli for functional face-area localizer)

We obtained functional face-area localizers for all participants (Fig. 1A). Stimuli for these localizers consisted of 42 black and white images of front-view faces of different identities and objects, respectively. Images of faces from different identities with a neutral expression were taken from the NimStim Face Stimulus Set (<http://www.macbrain.org/resources.htm>). Images of different objects were taken with a digital camera (Nikon D90, Nikon Corporation [Chiyoda, Tokyo, Japan]). All images were corrected for luminance (luminance corresponds to mean gray value 129.46 and contrast corresponds to standard deviation gray value 45.75) and were presented at a visual angle of $14 \times 10^\circ$ (Fig. 1A).

Procedure

All participants first took part in a training session (Fig. 1A) in which they learned to identify the voice, face, and name of three speakers. They then participated in a “psychophysical pilot experiment” that served to define the morph level of the face stimuli for the “cross-modal priming experiment” in the MRI scanner. The cross-modal priming experiment was followed by a standard functional face-area localizer (Fig. 1A).

Training

Before fMRI scanning, all participants were trained to identify the three male speakers. In the first part of the training (learning), participants watched audio-visual sam-

ples of each speaker’s face, voice, and additionally their written name (Anton, Bernd, and Carsten). Participants were asked to learn the association between voice, face, and name. We chose this particular training method because overt learning of faces and labeling a face with the corresponding name has been shown to be essential for categorical perception of morphed faces [Kikutani et al., 2008 2010]. Each video contained a five-word sentence lasting 2–3 s. There were 11 videos per speaker so that the total exposure to each speaker’s voice and face was about half a minute. Because there were three speakers, learning lasted about 2 min in total. The videos of the different speakers were presented fully randomized.

In the second part of the training (test), we tested whether participants could associate the correct name with the voice and the face of the three speakers. Responses were made via a three button response box with the labels A, B, C for the names of the three speakers (Anton, Bernd, and Carsten). The training test was designed to resemble the fMRI cross-modal priming experiment as closely as possible (see below). Each test trial comprised the presentation of an auditory-only voice (one word lasting ca. 1.6 s) followed by a muted video of the face (ca. 1.6 s). Participants were tested separately on their ability to identify the person’s face and voice. The trials were organized in six blocks of six trials each. In half of the blocks (voice task blocks), a colored fixation cross after the voice instructed participants to indicate via button press the corresponding name of the voice. In the other half of the blocks (face task blocks), the colored fixation cross occurred after the face stimulus and participants indicated the corresponding name of the face via button press. Participants received written feedback about correct, incorrect, and too slow responses (i.e., responses occurring more than 2 s after stimulus onset). Randomly interspersed in the test-part participants saw audio-visual videos of each speaker to refresh the association of face and voice (without colored fixation cross or task). These videos were presented in two blocks of six videos each (two videos per speaker).

The training, including learning and testing, took about 12 min in total (i.e., 2 min learning and about 10 min test). If a participant performed less than 80% in the test, the whole training was repeated. This was the case for 10 of the 16 participants; no participant needed more than two rounds of training. The training took place two days before scanning. Immediately before scanning the training was repeated once more for every participant to make sure that all participants remembered speakers’ faces and voices. All participants scored at least 80% in the training test immediately prior to fMRI-scanning.

Psychophysical pilot experiment

To select the morph levels for the fMRI cross-modal priming experiment, we conducted a psychophysical pilot experiment with all participants. The experiment had two

different sessions. In both sessions, participants were asked to report the name belonging to the faces of the three speakers via a three-button response box with the labels A, B, and C, corresponding to the initial letters of speakers' names.

In one session, participants performed this task on morphed faces with a morph continuum containing 11 morph levels (0–100%). This served to investigate the face morphing levels at which participants would perceive the face as belonging to the correct speaker. Previous studies have shown that this perception follows a step-like function and that different morphed faces are perceived as different in identity around the 50% morph level [Rotshtein et al., 2005]. The psychophysical pilot experiment showed that this was also the case for the stimulus set used for the present study. The polynomial regression for behavioral responses showed that the third-order component of the polynomial function differed significantly from zero (one-sample t -test: $t(15) = -18.0458$, $P < 10^{-10}$). The morph level corresponding to the perceptual change of identity (point of subjective equality, i.e., 50% of face responses) was within the morph-level range of 45–55% (one-sided one-sample t -test $> 45\%$: $t(15) = -3.1497$, $P = 0.0033$ and one-sided one-sample t -test $< 55\%$: $t(15) = 7.6744$, $P < 10^{-5}$).

The other psychophysical pilot experiment session served to ensure that there was also a step-like function in face perception when voices precede the faces. This was tested on a restricted set of five morph levels (i.e., 0, 33, 50, 67, and 100%). In each trial, the participants listened to a speaker's voice (two-word sentences, e.g., "Er sagt."; English: "He says.") and subsequently saw a face. The behavioral results confirmed that the voice prime did not abolish the step-like function found for face perception without voice primes (one-sample t -test: $t(15) = -15.4220$, $P < 10^{-8}$).

Based on these results, we chose to use four morph levels in the cross-modal priming experiment in the MRI scanner. These morph levels differed parametrically in physical properties with an equal distance along the morph continuum (0, 33, 67, and 100%); two morph levels were perceived as one identity and the two other morph levels were perceived as another identity (0/33% vs. 67/100%). We defined "mismatch of identity" as all trials in which the subsequent face contained only 0% or 33% of the preceding voice-identity. Vice versa, we defined "match of identity" as all trials in which the subsequent face contained 67% or 100% of the preceding voice-identity.

In addition, we collected discrimination performance along the continuum of the four different morph levels with a same-different categorization task; this served to test whether the morph levels were indeed perceived categorically [Beale and Keil, 1995]. We used the following design that differed slightly from procedures previously used [Beale and Keil, 1995; Rotshtein et al., 2005]. In each trial, participants had to categorize whether two subsequently represented face images belonged to the same or a different person. Responses were given via button-press. The first face image within each trial was always the 67% morph

condition (for each of the six morph combinations: A-B, A-C, B-A, B-C, C-A, and C-B). The second image within each trial was randomly selected from three possible conditions: (a) exactly the same image ("Identical": 67%), (b) the same identity, but at a different physical morph level ("Within": 100%), or (c) a different identity at a different physical morph level ("Between": 33%). Images were presented for 500 ms. Conditions (b) and (c) both represented images that changed by 33% in relation to the first image on each trial. The images within each trial were separated by 200 ms and trials were separated by 2.5 s. Before the main same-different categorization task, participants were familiarized with the three different identities and learned to associate each face (presented without morphing) with the corresponding name. There were five familiarization trials of each face-name identity. We collected data from nine participants (five females, mean age 29.22 years) who did not participate in the fMRI experiment. The experiment confirmed that perception of identity changed between the 67 to 33% condition (Between $>$ Within: 77.78% $>$ 17.59% responses "Different", paired t -test $t(8) = 8.374$, $P < 0.0001$), but not between the 67 and 100% condition (Within = Identical: 17.59% = 12.04%, responses "Different", paired t -test $t(8) = 1.333$, $P = 0.2191$), although the change in physical morph level was the same in both comparisons (33% from the first image).

fMRI—Cross-modal priming experiment

This experiment was an event related design (Fig. 1A). Each trial began with the presentation of a voice (two-word sentence in German, e.g., "Er baut." / English: "He builds.") for about 1,200 ms. The voice was followed by the presentation of a face image for 200 ms. Both stimulus types were separated by an interstimulus interval of 75 ms during which participants saw a fixation cross. The faces could have one of four different morph levels (0, 33, 67, and 100%). As revealed by the psychophysical pilot experiment, this resulted in trials where the voice identity and the perceived face identity matched (67 and 100%) or did not match (0 and 33%). At the same time, these morph levels differed parametrically in physical properties from the face belonging to the voice, that is, from 0% morph (no difference) to 100% morph (largest difference). Participants reported the identity of each face by pressing one of three response buttons (A, B, and C) corresponding to the initial letters of speakers' names ("face task," Fig. 1A). We used this forced-choice task instead of a same-different classification task because our aim was to investigate brain mechanisms for individual recognition and not for comparison of two stimuli. In 16% of the trials, participants were also asked to recognize voices ("voice task," Fig. 1A). These trials were excluded from the analyses. We did not include the voice-recognition task in all trials to prevent voice recognition task responses occurring at a similar time to the presentation of the face. The voice-recognition task was indicated by a colored cross appearing

subsequently to the voice presentation for 2,000 ms, but before the face. That way, during presentation of the voice participants did not know whether that trial would include a voice task or not. This ensured that participants were attending to the voices on all trials.

A trial lasted about 5,075 ms. If a voice-recognition task was included, the trial was 2,000 ms longer (7,075 ms). Trials were separated by a jittered interstimulus interval (range 1–12 s, mean 3,600 ms) during which a white fixation cross was shown. The experiment contained 90 trials for each of the four morph conditions (0, 33, 67, and 100%). In addition, 105 null events of 5,075 ms duration were included [Friston et al., 1999].

The trials were grouped into 15 blocks of 28 stimuli each, in which only two of the three speaker identities were presented. This was done for two reasons: First, it ensured that participants always made a categorical decision between the identities of two of the speakers. Second, it kept the probability of an occurrence for each identity $[P(i)]$ constant at both the global experiment level $[P(i) = 1/3]$ and local block level $[P(i) = 1/2]$.

The experiment was subdivided into three sessions of five blocks with short breaks (ca. 1 min) in-between the sessions. Total scanning time for the cross-modal priming experiment was about 46 min.

fMRI—Localizer of face-sensitive areas

We used a standard fMRI experiment to localize the OFA and FFA [Berman et al., 2010; Fox et al., 2009; Kanwisher et al., 1997]. In the experiment, images of faces and objects were presented in three blocks of each image category in alternating order. Each block consisted of 42 images. Each image was presented for 500 ms. There was no interstimulus interval between image presentations. Blocks were followed by a fixation cross that was shown for 18 s. Participants were instructed to look attentively at the images and the fixation cross. Total scanning time for the localizer experiment was 4 min.

Image Acquisition

Functional and structural T1-weighted images were acquired on a 3T Siemens Tim Trio MR scanner (Siemens Healthcare, Erlangen, Germany).

Functional MRI sequence

For the cross-modal priming experiment and the localizer of face-sensitive areas, a gradient-echo EPI (echo planar imaging) sequence was used (TE 30 ms, flip angle 90°, TR 2.79 s, 42 slices, whole brain coverage, acquisition bandwidth 116 kHz, 2-mm slice thickness, 1-mm interslice gap, in-plane resolution $3 \times 3 \text{ mm}^2$, ascending interleaved slice acquisition, 1,044 volumes in total for the cross-modal priming experiment, and 88 volumes in total for the localizer of face-sensitive areas). Geometric distortions were

characterized by a B0 field-map scan. The B0 field map was collected once per subject before each run. The field-map scan consisted of gradient-echo readout (24 echoes, interecho time 0.95 ms) with standard two-dimensional phase encoding. The B0 field was obtained by a linear fit to the unwrapped phases of all odd echoes.

Structural T1-MRI sequence

Structural images were acquired with a T1-weighted magnetization-prepared rapid gradient echo sequence (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 = 1,300 ms, TE = 3.93 ms, $\alpha = 10^\circ$, 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, The MathWorks, MA).

Functional MRI

Functional MRI data were analyzed with statistical parametric mapping (SPM8, Wellcome Trust Centre for Neuroimaging, UCL, UK, <http://www.fil.ion.ucl.ac.uk/spm>), using standard spatial preprocessing procedures (realignment and unwarp, coregistration of functional images to the participant's anatomical T1 image and normalization to Montreal Neurological Institute (MNI) standard stereotactic space, and smoothing with an isotropic Gaussian filter, 8 mm at full width at half maximum (FWHM) [Friston et al., 2007]. Statistical parametric maps were generated by modeling the evoked hemodynamic response as boxcars convolved with a synthetic hemodynamic response function in the context of the general linear model [Friston et al., 2007]. Hemodynamic responses were modeled at the onset of the presentation of the face image separated into the four different conditions (0, 33, 67, and 100%). In addition, we modeled the onset of voices in face-task and in voice-task trials as events of no interest. Therefore, the design matrix included five conditions.

Population-level inferences using Blood-oxygen-level dependent (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]. Results in Figures 3 and 4 were visualized with MRIcron (National Institute of Health, <http://www.nitrc.org/projects/mricron/>).

Activity analyses: Contrasts of interest, hypotheses, and regions of interest

Contrasts of interest and hypotheses. We computed two contrasts of interest. First, we computed a categorical

contrast (morph conditions $0\% + 33\% > 67\% + 100\%$). In this contrast, all conditions in which there was a mismatch between voice and face identity (morph levels 0% and 33% of the preceding voice-identity) were contrasted with conditions in which there was a match of identity (morph levels 67% , 100% of the preceding voice-identity). We used exactly the same original, unmorphed face images in the 0 and 100% conditions and exactly the same morphed face images in the 33 and 67% conditions; the crucial difference between the 0 and 100% and between the 33 and 67% conditions was the identity of the preceding voice prime that could either match or mismatch the identity of the following face image. Therefore, the differences in brain responses between conditions were caused only by the identity of the voice prime, not by differences in the face images. The contrast served to investigate whether face-sensitive areas respond with increased activity if there is a mismatch in voice and face identity (for reports of signal increases due to other mismatching stimuli see [Summerfield et al., 2008] and [Egner et al., 2010]). In previous priming studies with face-only stimuli, it has been shown that face-sensitive areas respond with higher activity to different identity of faces [Henson et al., 2000; Rotshtein et al., 2005]. Second, we computed a parametric contrast (morph conditions $0\% > 33\% > 67\% > 100\%$ with the contrast weights $[1 \ 1/3 \ -1/3 \ -1]$). This served to investigate which face-sensitive areas respond increasingly with increasing mismatch in physical properties from the face that belongs to the preceding voice identity.

We expected that the strength of the measured prediction errors depended on the level of certainty the participant had about the identity of the voice, that is, how fast the voices were recognized. To assure that the voice prime was effective to prime identity, participants were trained until they were able to recognize the identity of the voice prime. Therefore, performance correct in the voice-recognition task was at ceiling across participants. Consequently, we used reaction times as an indicator of certainty in voice-identity recognition. Models of decision making assume that evidence is accumulated over time until the decision is made [Gold and Shadlen, 2007; Smith and Ratcliff, 2004], and this assumption seems to be especially appropriate for dynamic stimuli, like the voice, which unfold information over time [Schweinberger et al., 1997a]. As an index for the participant's certainty about the identity of the voice, that is, how fast the voices were recognized; we used reaction time to voices in each individual participant averaged over all trials as a covariate on the second level. Ideally, we would have used individual trial voice recognition scores, however, this was not possible, because the voice task was only used in 16% of the trials to prevent voice recognition task responses occurring at a similar time to the presentation of the face. These trials were discarded from the analysis. The reaction time index was included as one regressor in the second level analysis with one value per participant. We expected that those participants that had a high certainty about voice

identity (expressed as fast reaction times) would also show the most prominent precision-weighted prediction error in visual-face areas during presentation of a mismatching face. This mechanism of predictive coding can also be reconciled with the interpretation of repetition suppression [Friston, 2005; Garrido et al., 2008; Todorovic and de Lange, 2012].

Throughout the text, we will refer to the two contrasts of interests including the correlation as “identity contrast” and “physical-distance contrast.” To test for significant differences between the two contrasts, we extracted the eigenvariate for each contrast from independently localized face-sensitive areas and compared depending correlations using a standard statistical approach [Steiger, 1980].

Prompted by the helpful suggestions of a reviewer, we performed two additional correlation analyses. First, to rule out that the obtained results are caused by general differences between “slow” and “fast” participants and to make sure that the resulting correlations were specific for reaction times in voice recognition we repeated the analysis with reaction times in the voice task that were normalized for overall reaction times (by subtracting reaction times in face recognition from reaction times in voice recognition for each participant).

In a second additional analysis, we used reaction times in the face-recognition task as parametric regressors at the first level (one regressor for each of the four face morph conditions with the face reaction times for each single trial).

Regions of interest. Regions of interest (ROI) were the three face-sensitive areas in the right hemisphere that were implied in face recognition: OFA, FFA, and aTL [Garrido et al., 2009; Grill-Spector et al., 2004; Kriegeskorte et al., 2007; Nestor et al., 2011; Pitcher et al., 2011; Pourtois et al., 2005; Rajimehr et al., 2009; Rossion, 2008; Rotshtein et al., 2005]. We localized the OFA and FFA with the standard contrast visual “face stimuli vs. object stimuli” [Berman et al., 2010; Fox et al., 2009; Kanwisher et al., 1997]. The FFA was localized at $x = 42$, $y = -49$, $z = -20$ (63 voxels) and the OFA at $x = 39$, $y = -76$, $z = -20$ (10 voxels). The standard localizer does not provide information about the location of the face-sensitive area in the aTL [Kanwisher et al., 1997]. To define the aTL region of interest, we therefore used a sphere (4 mm radius) around a previously published coordinate [Garrido et al., 2009] resulting in a region of interest for the aTL at the coordinates $x = 39$, $y = -4$, $z = -35$. This region of interest contained also the coordinates from another previously published study ($x = 39$, $y = -4$, $z = -32$, when transformed to MNI space, [Rajimehr et al., 2009]).

Connectivity: Psychophysiological interaction analyses and ROI

Psychophysiological interactions. With fMRI, following the predictive coding account, we expected an increased functional connectivity during the presentation of a

TABLE I. Behavioral results in voice and face recognition tasks ($n = 16$)

Task: Mean (std)	Voice task	Face task 100% morph	Face task 67% morph	Face task 33% morph	Face task 0% morph
Reaction times in ms	772.13 (133.12)	703.28 (63.4)	770.96 (77.21)	727.95 (94.01)	644.68 (61.7)
Correct responses in %	94.625 (6.407)	98 (1.9)	94.63 (3.7)	96.19 (2.56)	98.69 (2.21)

mismatching face following a voice prime, compared to a matching face. First, we specifically tested the cross-modal coupling of FFA and voice-sensitive areas in middle/anterior STS because these areas have been shown to be structurally connected [Blank et al., 2011]. Second, we tested intramodal coupling between FFA and OFA and between FFA and aTL, because these areas have also been shown to be structurally connected [Gschwind et al., 2012; Thomas et al., 2009]. To investigate functional connectivity between FFA and middle/anterior STS voice-sensitive areas and between face-sensitive areas, we conducted two sets of psychophysiological interaction (PPI) analyses [Friston et al., 1997]. Regressors for these analyses included the psychological variable, the time course of the region of interest, and the psychophysiological variable. For the first set of PPI analyses, the psychological variable was the categorical “identity contrast” ($0\% + 33\% > 67\% + 100\%$ with the contrast weights $[1\ 1\ -1\ -1]$). For the second set of PPI analyses, the psychological variable was the parametric “physical-distance contrast” ($0\% > 33\% > 67\% > 100\%$ with the contrast weights $[1\ 1/3\ -1/3\ -1]$). Each set of PPI analyses contained three separate analyses that differed in the region of interest from which the eigenvariables for the time-course regressor were extracted (i.e., right OFA, right FFA, right aTL; see detailed description in the section below: Regions of interest). These regions were defined by the categorical “identity contrast” for the first set of PPI analyses and the parametric “physical-distance contrast” for the second set of PPI analyses. The psychophysiological variable was created using routines implemented in SPM8. PPI regressor, psychological variable, and first eigenvariate were entered in a design matrix at the single-subject level. Population-level inferences about BOLD signal changes were based on a random-effects model that estimated the second-level statistic at each voxel using a one-sample t -test. Each second level design matrix included a covariate with each individual participant’s response times for voices.

ROIs for the PPI analyses. We used ROIs to test for significant effects of the PPI analysis. We defined these ROIs in the following way: we expected and tested functional connectivity (i) between FFA and the voice-identity sensitive STS [Blank et al., 2011]; (ii) between FFA and the OFA [Gschwind et al., 2012]; (iii) and between FFA and the

face-sensitive areas in the aTL. The voice-sensitive area was determined as a sphere of 4 mm radius around a coordinate $[x = 60, y = 0, z = -15]$ that has been identified as auditory voice-sensitive area in the anterior/middle STS previously [von Kriegstein and Giraud, 2004]. The visual areas of interest were defined by the activity clusters obtained with the “identity contrast” and the “physical-distance contrast” in the OFA, FFA, and aTL.

For PPI in event-related designs, the deconvolution process implemented by SPM is particularly sensitive to noise and HRF inaccuracies. In addition, implementing PPI with seeds that have a response to the contrast of interest can lead to regressor collinearity issues and a lack of power because the physiological and psychological variables may already be correlated, in which case the PPI regressor could also be correlated with both of these. These constraints make it more difficult to find significant PPI results in our event-related design.

Statistical significance criterion

Functional activity and connectivity results were considered significant if they were present at $P < 0.05$ family-wise error (FWE) corrected for the ROI. In Figures 3 and 4, for display purposes only, results were displayed at $P < 0.01$, masked by the ROI.

RESULTS

Behavioral Results

Participants achieved high performances in both voice-identity recognition ($M = 94.625$, $STD = 6.407$, $n = 16$) and face-identity recognition ($M = 98.625$, $STD = 1.628$, $n = 16$) in the cross-modal priming experiment (Table I). The behavioral performance of this experiment confirmed that the morphed faces were perceived in a categorical manner: The third-order component of a polynomial expansion differed significantly from zero (one-sample t -test) showing that faces with 0 and 33% morph were recognized as one identity and faces with 67 and 100% morph were recognized as the other identity (Fig. 2A). The third-order component of a polynomial expansion differed significantly from zero (one-sample t -test: $t(15) = -45.767867$, $P < 10^{-16}$).

Analysis of correct responses and reaction times (Table I, Fig. 2B) showed that face recognition was easier and temporally facilitated when the preceding voice was of the same identity as the subsequently presented face (main

Figure 2.

Behavioral Results. A. Behavioral responses of individual participants confirm that morphed faces were perceived in a categorical manner: Faces with 0 and 33% morph were recognized as one identity, whereas faces with 67 and 100% morph were recognized as the other identity. “Mismatch identity” conditions were defined as trials in which the subsequent face contained only 0% or 33% of the preceding voiceidentity. Vice versa, “match identity” condi-

tions were defined as trials in which the subsequent face contained 67% or 100% of the preceding voice-identity. **B.** Reaction times were facilitated in trials when identity of face and voice matched in identity (light gray bars) compared to when they mismatched (dark gray bars). Error bars indicate standard error of the mean. Stars indicate the significant main effect of the factor “mismatch vs. match of identity” ($F(1,15) = 57.472$, $P < 0.001$).

effect of the factor “mismatch vs. match of identity”; correct response: $F(1,15) = 7.642$, $P = 0.0145$; response time: $F(1,15) = 57.472$, $P < 0.001$, Table I, Fig. 2B). In addition, morphed faces were recognized with more difficulty and at a slower rate than normal faces (main effect of the factor “normal vs. morphed”; correct response: $F(1,15) = 15.696$; response time: $F(1,15) = 42.357$, $P < 0.001$, $P = 0.0013$, Table I, Fig. 2B). There was no interaction of the factors “mismatch vs. match” and “normal vs. morphed” (correct response: $F(1,15) = 0.547$; response time: $F(1,15) = 2.089$, $P = 0.1689$, $P = 0.4709$).

fMRI

Voices activate representations of physical properties and conceptual identity of faces in distinct face-sensitive areas

The activity analysis without including the covariate (RT in voice-identity recognition) did not reveal significant results within our ROI. We next report the results using reaction time in voice-identity recognition as a covariate at the second level.

We first addressed the question whether recognizing voices leads to representations of face identity in any of the three face-sensitive areas (Fig. 3A). To answer this question, we used the categorical “identity contrast” ($0\% + 33\% > 67\% + 100\%$ with reaction times in voice-identity recognition as covariate). For this contrast, a significant correlation with reaction times in voice-identity recognition was present in OFA, in FFA, and in

aTL (Table II, Fig. 3B–D). We interpret this mismatch response as a representation of prediction errors if the face does not match the voice in identity.

We then tested whether recognizing voices with high certainty leads to representations of physical properties in any of the three face-sensitive areas. To test this, we used the parametric “physical-distance contrast” ($0\% > 33\% > 67\% > 100\%$ with reaction times in voice-identity recognition as covariate) (Fig. 1B, light blue). For this contrast, a significant correlation with reaction times in voice-identity recognition was present only in right OFA (Table II, Fig. 3B). We interpret this response as a representation of increasing prediction errors with more physical distance from the face that belongs to the voice prime.

In a next step, we tested to what extent the voice elicits specialized responses to face identity or physical properties in the different face-sensitive areas. To do this, we tested for significant differences between the correlation of the parametric “physical-distance contrast” and the categorical “identity contrast.” In aTL, the correlation with the “identity contrast” was significantly stronger than correlation with the “physical-distance contrast” ($t(13) = 2.4037$, $P = 0.0319$, Fig. 3D). In contrast, in OFA and FFA, there was no significant difference between the two correlations ($t(13) = -0.6617$, $P = 0.5197$ and $t(13) = 1.5595$, $P = 0.1429$, Fig. 3B,C, respectively).

The analysis with the normalized reaction times in voice recognition (i.e., reaction time to voices minus to reaction time to faces) provided similar results as the first analysis using the original voice recognition reaction times. The correlation of the “identity contrast” with the normalized

Figure 3.

Effects of voice primes in face-sensitive regions. **A.** ROI displayed on rendered Colin brain image (OFA (cyan) and FFA (light blue) defined by activity in the localizer contrast, and aTL (dark blue) defined by published coordinates (see Methods). **B–D.** Results for the correlation with the categorical “identity contrast” and the parametric “physical-distance contrast” (i.e., identity contrast: $(0\% + 33\%) > (67\% + 100\%)$ and physical-distance contrast: $(0\% > 33\% > 67\% > 100\%)$, both correlated with the reaction time to voices). **B.** In right OFA, there were significant correlations with

both the “identity contrast” and the “physical-distance contrast” and the effects did not differ significantly from each other. **C.** In right FFA (light blue), there was a significant correlation with the “identity contrast” and not for the “physical-distance contrast,” but the effects did not differ significantly from each other. **D.** In right aTL (dark blue), there was a significant correlation with the “identity contrast” and not for the “physical-distance contrast”. Specifically, the correlation with the “identity contrast” was significantly stronger than the one for the “physical-distance contrast”.

reaction times in voice recognition was significant in the FFA ($x = 45$, $y = -52$, $z = -26$, $P = 0.034$, FWE corrected for FFA), in the OFA ($x = 36$, $y = -73$, $z = -20$, $P = 0.005$, FWE corrected for OFA), and in the aTL ($x = 36$, $y = -4$, $z = -35$, $P = 0.003$, FWE corrected for aTL).

The correlation of the “physical-distance contrast” with the normalized reaction times in voice recognition was significant in the OFA ($x = 39$, $y = -73$, $z = -20$, $P = 0.011$, FWE corrected for OFA), and for this analysis also in the aTL ($x = 36$, $y = -4$, $z = -35$, $P = 0.043$, FWE corrected for aTL). There was a trend to significance in the FFA ($x = 45$, $y = -37$, $z = -11$, $P = 0.057$, FWE corrected for FFA).

The comparison of the correlations between “physical-distance contrast” and the “identity contrast” confirmed the findings of the first analysis: only in the face-sensitive aTL, the correlation of reaction times in voice recognition with

the “identity contrast” was significantly stronger than with the “physical distance contrast” ($t(13) = 2.4119$, $P = 0.031$), whereas the correlations did not differ in OFA ($t(13) = -0.3321$, $P = 0.7451$), and FFA ($t(13) = 1.3301$, $P = 0.2063$).

The analysis with reaction times in the face-recognition task as parametric regressors at the first level did not reveal any significant results within our ROI (even at a lenient threshold of $P < 0.1$) and also not at a level of $P < 0.05$ FWE corrected for the whole brain.

Functional connectivity between face- and voice-sensitive areas increases if the face identity and physical properties do not match the voice

Using functional connectivity, we addressed the question whether mismatch of identity or physical properties between

TABLE II. Overview of fMRI results for three face-sensitive regions (OFA, FFA, and aTL)

Effects of voice primes in face-sensitive regions				
Face-sensitive region	MNI coordinates (x, y, z)	Identity contrast	Physical-distance contrast	Identity > physical-distance
Right OFA	39, -73, -20	$P = 0.008$, FWE-corrected for OFA $r = -0.592$	$P = 0.004$, FWE-corrected for OFA $r = -0.6311$	n.s. $P = 0.5197$, $t(13) = -0.6617$
Right FFA	45, -52, -26	$P = 0.029$, FWE-corrected for FFA $r = -0.542$	n.s. $P = 0.061$, FWE-corrected for FFA $r = -0.4339$	n.s. $P = 0.1429$, $t(13) = 1.5595$
Right aTL	36, -7, -35	$P = 0.013$, FWE-corrected [Garrido et al., 2009] $r = -0.606$	n.s. $P > 0.1$, FWE-corrected [Garrido et al., 2009], $P > 0.01$, uncorrected $r = -0.2932$	$P = 0.0319$, $t(13) = 2.4037$

x , y , and z indicate peak coordinates in MNI, n.s. indicates nonsignificance at $P > 0.05$.

voice and face also leads to an increased functional connectivity between (i) middle/anterior STS voice-sensitive areas and FFA and (ii) FFA, OFA, and aTL. Such a connectivity result would suggest that FFA exchanges information cross-modally with the voice-sensitive middle/anterior STS and intramodally with face-sensitive OFA and aTL.

Cross-modal functional connectivity. We found that functional connectivity between FFA and middle/anterior STS correlated with voice recognition reaction times for both types of mismatch (i.e., the mismatch of identity and the mismatch of physical properties) in contrast to a match of voice and face. This was measured with the “identity contrast” and the “physical-distance contrast” (Fig. 4). The

Figure 4.

Effects of voice primes on functional connectivity. Functional connectivity between FFA (light blue) and voice-sensitive STS (red) and between FFA and OFA (cyan) was enhanced for both the “identity contrast” and the “physical-distance contrast” (i.e., identity contrast: $(0\% + 33\%) > (67\% + 100\%)$ and physical-distance contrast: $(0\% > 33\% > 67\% > 100\%)$, both correlated with the reaction time to voices).

statistical maximum in the middle/anterior STS was at $x = 63$, $y = -1$, $z = -14$ (for both analyses, i.e., for the “identity contrast” and the “physical-distance contrast”: $P = 0.022$, FWE-corrected for the ROI in the voice-sensitive middle/anterior STS [von Kriegstein and Giraud, 2004], eigenvariate extracted from FFA). The same coordinate was obtained for the correlation with the normalized reaction time in voice recognition (for the “identity contrast”: $P = 0.019$ and “physical-distance contrast”: $P = 0.035$, FWE-corrected). We interpret the increased functional connectivity during the mismatch of face identity and facial properties as an exchange of predictions and/or prediction errors about both face identity and physical distance with voice-sensitive middle/anterior STS. The correlation with functional connectivity modeled with the “identity contrast” and the “physical-distance contrast” did not differ significantly from each other, indicating that functional coupling between FFA and STS was similar for the mismatch of both identity and physical distance ($t(13) = -0.2904$, $P = 0.7761$, and also with normalized reaction times: $t(13) = -0.4778$, $P = 0.6407$). This fits the activity findings where the correlation with responses in the FFA did not differ significantly for the “identity contrast” and the “physical-distance contrast” suggesting that during both conditions information might be communicated between the voice-sensitive area and the FFA.

Intramodal functional connectivity. In addition, we also analyzed the functional coupling between the face-sensitive regions to test whether it depends on a mismatch in face identity or physical facial properties. There was a correlation of intramodal functional connectivity between face-sensitive areas with reaction times in voice-identity recognition (Fig. 4): The functional connectivity between FFA and OFA correlated with reaction times in voice recognition (for both the “identity and physical-distance contrast”: $x = 39$, $y = -73$, $z = -20$, $P = 0.008$ and $P = 0.004$, FWE-corrected for OFA; eigenvariate extracted from FFA). There was a trend toward significance showing that the connectivity between OFA and FFA was stronger during physical-distance mismatch than identity mismatch,

$t(13) = -2.1338$, $P = 0.0525$), which mirrors the activity findings for this region (Fig. 4B).

Again, we obtained the same results when we used the normalized reaction time covariate (for the “identity contrast”: $x = 39$, $y = -73$, $z = -20$, $P = 0.014$ and for the “physical-distance contrast”: $x = 36$, $y = -73$, $z = -23$, $P = 0.003$, FWE-corrected for OFA; eigenvariate extracted from FFA). These correlations with functional connectivity modeled with the “identity contrast” and the “physical-distance contrast” did not differ significantly from each other ($t(13) = -0.5551$, $P = 0.5882$).

There were no significant results in our ROIs and at $P < 0.05$ FWE-corrected for the whole brain without the covariate.

DISCUSSION

In the present study, we tested whether voice recognition activates representations of physical properties of faces or conceptual identity of faces in face-sensitive regions. We used fMRI and a voice-face priming design in which familiar voices were followed by morphed faces that matched or mismatched with respect to identity or physical properties. We found that functional activity and connectivity correlated with reaction times to voices. Specifically, activity due to a mismatch of voice and face identity correlated with reaction times to voices in OFA, in FFA, and in aTL, while activity due to a mismatch of physical properties between voice and face correlated with reaction times to voices only in OFA. These findings indicate that hearing a voice elicits representations of physical properties as well as the identity of a person in face-sensitive areas. The representations seemed to be specific to the brain area: (i) The OFA represented information about both physical properties and identity, (ii) the FFA represented identity, and (iii) the aTL represented identity information to a greater extent than information about physical properties. In addition, there was increased cross-modal functional connectivity between face- and voice-sensitive areas (i.e., FFA and middle/anterior STS) as well as intramodal connectivity between face-sensitive areas (i.e., FFA and OFA) when the face did not match the voice in identity or physical properties. These results delineate the neural mechanism for cross-modal recognition of individuals and suggest that this process relies on exchanging information about both identities of a face and physical facial properties across sensory modalities. We speculate that during recognition of familiar voices the face-sensitive areas are tuned to physical features and to conceptual identity of faces.

The current findings advance previous work on voice-face integration [Blank et al., 2011; Campanella and Belin, 2007; Schweinberger et al., 2011; von Kriegstein and Giraud, 2006; von Kriegstein et al., 2008], because they characterize what type of information might be exchanged between voice and face areas. Furthermore, they suggest

for the first time that a familiar voice can alter the neural responses to faces in a very specific way: physical properties as well as identity information seem to be shared between specialized voice- and face-sensitive regions and all face-sensitive areas are involved in this cross-modal process.

However, we were not able to find evidence for cross-modal activation in face-sensitive regions without including the reaction times to voices as a covariate to the analysis. In addition, we were not able to find a correlation with reaction times in the face-recognition task. We presented a set of three different speaker identities in the current design, so that the presentation of a mismatching face might have been overall not very unexpected. We speculate that this explains why the mismatch signal was very small. Nevertheless, the results contribute to the question about which representations are activated by voices in different face-sensitive brain regions during cross-modal recognition. The effect of the covariate suggests that activity during presentation of a mismatching face was higher in subjects with a better/faster performance in voice-identity recognition than in subjects with worse/slower recognition performance. Future studies manipulating levels of recognition in a within-subject design would help to further strengthen this finding. Morphing the voice prime [Latinus et al., 2011; Zaske et al., 2010] might provide another promising approach to manipulate and test the strength of cross-modal predictions.

Our results indicate that the exchange of cross-modal information becomes more specific for identity in more anterior face-sensitive areas. This integrates well with the ongoing debate about where representations of face identity, in contrast to physical face properties, are processed within the face recognition network (for review [Atkinson and Adolphs, 2011; Haxby et al., 2000; Tsao and Livingstone, 2008; Von Der Heide et al., 2013]). We found evidence that cross-modal representations in the aTL are relatively specific for identity. This identity-specificity of the aTL is in line with several studies investigating identity representation of faces [Kriegeskorte et al., 2007; Nasr and Tootell, 2012; Rajimehr et al., 2009; Rotshtein et al., 2005; Tanji et al., 2012]. Further support for the role of the aTL in face-identity processing is also provided by studies on developmental prosopagnosia, a selective impairment in face-identity recognition. In developmental prosopagnosic participants, studies showed reduced gray matter volume within aTL [Behrmann et al., 2007; Garrido et al., 2009], reduced white matter connectivity between FFA and aTL [Thomas et al., 2009], and reduced functional connectivity between FFA and aTL [Avidan et al., 2014; von Kriegstein et al., 2006].

For the FFA, our results suggest that voices activate primarily representations of face identity, but these identity-specific representations were not significantly stronger than those of physical properties. This is in agreement with previous studies investigating identity and physical property representation of faces. These studies showed

that the FFA is involved in the processing of face identity, but to a certain extent also processes physical facial properties [Eger et al., 2004; Grill-Spector et al., 2004; Kriegeskorte et al., 2007; Nestor et al., 2011; Pourtois et al., 2005; Rotshtein et al., 2005; Xu et al., 2009].

The OFA has traditionally been considered to process physical facial properties at an early stage during face perception [Haxby et al., 2000; Pitcher et al., 2007, 2011] and the relation between facial properties [Rotshtein et al., 2007a]. In contrast to a strict hierarchical view of only early processing in OFA, there is evidence that OFA receives input from FFA [Jiang et al., 2011] and other areas in extrastriate cortex [Kim et al., 2006; Rotshtein et al., 2007b] and is also involved in processing face identity [Atkinson and Adolphs, 2011; Rossion, 2008]. Recently, multivariate pattern analysis revealed orientation-tolerant representations of face identity not only in aTL but also in OFA [Anzellotti et al., 2014]. Our results also support the involvement of OFA in processing both physical distance of facial properties and face identity. Receptive fields in OFA are supposed to be smaller than those in FFA and reactivation of more detailed processing in OFA could be essential for individualization of faces [Atkinson and Adolphs, 2011; Rossion, 2008]. In this view, the FFA may not be sufficient for face recognition without interaction with the OFA [Rossion et al., 2003; Steeves et al., 2006]. In relation to cross-modal individual recognition, it could mean that the FFA received cross-modal information from voice-sensitive regions during voice recognition and further communicated with the OFA to obtain more fine-grained information about facial properties. The correlations in the OFA and FFA in the present study have, however, to be interpreted with caution. The categorical and the parametric contrasts are correlated and it is therefore challenging to distinguish effects of physical similarity and identity when there is no significant difference between them. In principle, it would be possible, that in OFA and FFA only the identity or the physical properties are driving the correlation.

The potential transfer of identity and physical properties between voice- and face-sensitive areas is in line with two behavioral results in the literature. First, faces are recognized faster if they have been primed with a matching voice identity [Ellis et al., 1997] and vice versa for voice recognition [Schweinberger et al., 1997b]. The behavioral results of the present study replicated these effects, that is, voice primes facilitated recognition of matching faces (Fig. 2B). The proposed mechanism of information transfer between face- and voice-sensitive regions might explain this finding: voices activate the corresponding face representation in face-sensitive regions and the following face image could be compared against this prior activation at multiple levels of the face-recognition system and, therefore, face recognition is accelerated. Second, previous studies showed shifts in identity perception after cross-modal adaptation. Specifically, it has been found that after adaptation to face A, a voice morphed between speaker A and B is perceived as speaker B [Hills et al., 2010; Ziske et al.,

2010]. We speculate that after adaptation to face A, a voice morphed between speaker A and B causes a strong mismatch signal because the cross-modally activated face template of this morphed voice deviates from the previously adapted face image. Such a detected mismatch might shift perception to speaker B.

Although cross-modal individual recognition is important in many species [Kondo et al., 2012; Proops et al., 2009; Sai, 2005; Sliwa et al., 2011], the underlying computational mechanism is currently unknown. The present study was motivated by a predictive coding framework and the findings are compatible with such an account [Friston, 2005; Friston and Kiebel, 2009; Rao and Ballard, 1999]: we speculate that a precise representation of voice identity (indicated in the present study by fast reaction times) led to strong cross-modal predictions in face-sensitive areas; this in turn caused stronger precision-weighted prediction errors during conditions of mismatching voice and face. The increased functional connectivity of FFA and voice-sensitive middle/anterior STS during recognition of a mismatching face potentially reflects that prediction errors were sent back from face- to voice-sensitive areas. This would imply that (1) during recognition of a voice, identity, and physical properties of the corresponding face are encoded by FFA and that (2) during recognition of an unpredicted face, prediction errors are communicated back from FFA to voice-sensitive middle/anterior STS. An alternative interpretation of the correlation between reaction time in the voice-identity recognition task and the strength of the measured signal during presentation of a mismatching face could also be the framework of (a release in) repetition suppression during priming of voices and faces [Henson et al., 2000; Rotshtein et al., 2005]. However, recently this repetition suppression framework has been interpreted as an expression of a predictive coding mechanism [Friston, 2005; Garrido et al., 2008; Todorovic and de Lange, 2012].

The finding of intramodal functional connectivity between face-sensitive areas is in line with previous studies showing functional and structural connectivity between these regions [Gschwind et al., 2012; Mechelli et al., 2004; Thomas et al., 2009]. We interpret the intramodal coupling between OFA and FFA as reflecting an exchange of prediction errors resulting from the mismatch of expected physical properties and face identity. These interpretations should, however, be seen with caution, as a PPI analysis does not allow conclusions about the direction of the functional connectivity.

There was no increased functional connectivity to the face-sensitive area in the aTL from either OFA or FFA. Although we were able to find functional activity results in the aTL, the negative finding of functional connectivity between aTL and FFA could be due to the fact that investigation of the aTL with fMRI is challenging owing to susceptibility artifacts in aTLs [Devlin et al., 2000].

The transfer of information between visual and auditory regions during person-identity recognition is also in line

with cross-modal processing during recognition of object identity. Several fMRI studies have shown that stimulus-specific patterns of activity can be decoded across visual and auditory modalities [de Haas et al., 2013; Hsieh et al., 2012; Meyer et al., 2010].

In summary, our results suggest that the human brain uses voice information to activate both representations of identity and physical properties of faces in face-sensitive areas (OFA, FFA, and aTL), for example, when somebody calls out “Hello” from behind [Sheffert and Olson, 2004; von Kriegstein et al., 2008]. We speculate that exchange of information between face- and voice-sensitive areas is important to accomplish fast and robust recognition of individuals during human interactions and is an evolutionary conserved process that might also exist in the many other species that rely on vocal and facial identity recognition in their social interactions [Adachi et al., 2007; Kondo et al., 2012; Proops et al., 2009; Sidtis and Kreiman, 2012; Sliwa et al., 2011].

ACKNOWLEDGMENT

The authors thank Sonja Schall for very helpful comments on earlier versions of this manuscript. This work was funded by a Max Planck Research Group grant to KVK.

REFERENCES

- Adachi I, Kuwahata H, Fujita K (2007): Dogs recall their owner’s face upon hearing the owner’s voice. *Anim Cogn* 10:17–21.
- Andics A, McQueen JM, Petersson KM, Gal V, Rudas G, Vidnyanszky Z (2010): Neural mechanisms for voice recognition. *Neuroimage* 52:1528–1540.
- Anzellotti S, Fairhall SL, Caramazza A (2014): Decoding representations of face identity that are tolerant to rotation. *Cereb Cortex* 24:1988–1995.
- Atkinson AP, Adolphs R (2011): The neuropsychology of face perception: Beyond simple dissociations and functional selectivity. *Philos Trans R Soc B Biol Sci* 366:1726–1738.
- Avidan G, Tanzer M, Hady-Bouziene F, Liu N, Ungerleider LG, Behrmann M (2014): Selective dissociation between core and extended regions of the face processing network in congenital prosopagnosia. *Cereb Cortex* 26:1565–1578.
- Beale JM, Keil FC (1995): Categorical effects in the perception of faces. *Cognition* 57:217–239.
- Beauchemin M, Gonzalez-Frankenberger B, Tremblay J, Vannasing P, Martinez-Montes E, Belin P, Beland R, Francoeur D, Carceller AM, Wallois F, Lassonde M (2011): Mother and stranger: An electrophysiological study of voice processing in newborns. *Cereb Cortex* 21:1705–1711.
- Behrmann M, Avidan G, Gao F, Black S (2007): Structural imaging reveals anatomical alterations in inferotemporal cortex in congenital prosopagnosia. *Cereb Cortex* 17:2354–2363.
- Belin P, Zatorre RJ (2003): Adaptation to speaker’s voice in right anterior temporal lobe. *Neuroreport* 14:2105–2109.
- Belin P, Bestelmeyer PEG, Latinus M, Watson R (2011): Understanding voice perception. *Br J Psychol* 102:711–725.
- 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.
- Blank H, Anwander A, von Kriegstein K (2011): Direct structural connections between voice- and face-recognition areas. *J Neurosci* 31:12906–12915.
- Bruce V, Young A (1986): Understanding face recognition. *Br J Psychol* 77:305–327.
- Campanella S, Belin P (2007): Integrating face and voice in person perception. *Trends Cogn Sci* 11:535–543.
- de Haas B, Schwarzkopf DS, Ungerleider LG, Rees G (2013): Auditory modulation of visual stimulus encoding in human retinotopic cortex. *Neuroimage* 70:258–267.
- DeCasper AJ, Fifer WP (1980): Of human bonding: Newborns prefer their mothers’ voices. *Science* 208:1174–1176.
- Devlin JT, Russell RP, Davis MH, Price CJ, Wilson J, Moss HE, Matthews PM, Tyler LK (2000): Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task. *Neuroimage* 11:589–600.
- Eger E, Schyns PG, Kleinschmidt A (2004): Scale invariant adaptation in fusiform face-responsive regions. *Neuroimage* 22:232–242.
- Egner T, Monti JM, Summerfield C (2010): Expectation and surprise determine neural population responses in the ventral visual stream. *J Neurosci* 30:16601–16608.
- Ellis HD, Jones DM, Mosdell N (1997): Intra- and inter-modal repetition priming of familiar faces and voices. *Br J Psychol* 88: 143–156.
- Ethofer T, Bartscher J, Gschwind M, Kreifelts B, Wildgruber D, Vuilleumier P (2012): Emotional voice areas: Anatomical location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cereb Cortex* 22:191–200.
- Focke J, Holig C, Best A, Roder B (2011): Crossmodal interaction of facial and vocal person identity information: An event-related potential study. *Brain Res* 1385:229–245.
- 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.
- Freiwald WA, Tsao DY, Livingstone MS (2009): A face feature space in the macaque temporal lobe. *Nat Neurosci* 12:1187–1196.
- Friston K (2005): A theory of cortical responses. *Philos Trans R Soc London [Biol]* 360:815–836.
- Friston K, Kiebel S (2009): Predictive coding under the free-energy principle. *Philos Trans R Soc B Biol Sci* 364:1211–1221.
- Friston K, Ashburner A, Kiebel S, Nichols T, Penny W, editors (2007): *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.
- Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997): Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* 6:218–229.
- Friston KJ, Zarahn E, Josephs O, Henson RNA, Dale AM (1999): Stochastic designs in event-related fMRI. *Neuroimage* 10:607–619.
- Garrido L, Furl N, Draganski B, Weiskopf N, Stevens J, Tan GC, Driver J, Dolan RJ, Duchaine B (2009): Voxel-based morphometry reveals reduced grey matter volume in the temporal cortex of developmental prosopagnosics. *Brain* 132(Pt 12):3443–3455.
- Garrido MI, Friston KJ, Kiebel SJ, Stephan KE, Baldeweg T, Kilner JM (2008): The functional anatomy of the MMN: A DCM study of the roving paradigm. *Neuroimage* 42:936–944.
- Gold JI, Shadlen MN (2007): The neural basis of decision making. *Annu Rev Neurosci* 30:535–574.

- Grill-Spector K, Knouf N, Kanwisher N (2004): The fusiform face area subserves face perception, not generic within-category identification. *Nat Neurosci* 7:555–562.
- Gschwind M, Pourtois G, Schwartz S, de Ville DV, Vuilleumier P (2012): White-matter connectivity between face-responsive regions in the human brain. *Cereb Cortex* 22:1564–1576.
- Haxby JV, Hoffman EA, Gobbini MI (2000): The distributed human neural system for face perception. *Trends Cogn Sci* 4: 223–233.
- Henson R, Shallice T, Dolan R (2000): Neuroimaging evidence for dissociable forms of repetition priming. *Science* 287:1269–1272.
- Hills PJ, Elward RL, Lewis MB (2010): Cross-modal face identity aftereffects and their relation to priming. *J Exp Psychol Hum Percept Perform* 36:876–891.
- Hsieh PJ, Colas JT, Kanwisher N (2012): Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. *J Neurophysiol* 107:3428–3432.
- Jiang F, Dricot L, Weber J, Righi G, Tarr MJ, Goebel R, Rossion B (2011): Face categorization in visual scenes may start in a higher order area of the right fusiform gyrus: Evidence from dynamic visual stimulation in neuroimaging. *J Neurophysiol* 106:2720–2736.
- Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E (2003): “Putting the face to the voice”: Matching identity across modality. *Curr Biol* 13:1709–1714.
- 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.
- Kikutani M, Roberson D, Hanley JR (2008): What’s in the name? Categorical perception for unfamiliar faces can occur through labeling. *Psychon Bull Rev* 15:787–794.
- Kikutani M, Roberson D, Hanley JR (2010): Categorical perception for unfamiliar faces: The effect of covert and overt face learning. *Psychol Sci* 21:865–872.
- Kim M, Ducros M, Carlson T, Ronen I, He S, Ugurbil K, Kim DS (2006): Anatomical correlates of the functional organization in the human occipitotemporal cortex. *Magn Reson Imaging* 24: 583–590.
- Kondo N, Izawa E-I, Watanabe S (2012): Crows cross-modally recognize group members but not non-group members. *Proc R Soc B Biol Sci* 279:1937–1942.
- Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007): Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci USA* 104:20600–20605.
- Latinus M, Belin P (2011): Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in psychology* 2: 175.
- Mavica LW, Barenholtz E (2013): Matching voice and face identity from static images. *J Exp Psychol Hum Percept Perform* 39: 307–312.
- Mechelli A, Price CJ, Friston KJ, Ishai A (2004): Where bottom-up meets top-down: Neuronal interactions during perception and imagery. *Cereb Cortex* 14:1256–1265.
- Meyer K, Kaplan JT, Essex R, Webber C, Damasio H, Damasio A (2010): Predicting visual stimuli on the basis of activity in auditory cortices. *Nat Neurosci* 13:667–668.
- Mumford D (1992): On the computational architecture of the neocortex. 2. The role of corticocortical loops. *Biol Cybern* 66:241–251.
- Nasr S, Tootell RBH (2012): Role of fusiform and anterior temporal cortical areas in facial recognition. *Neuroimage* 63:1743–1753.
- Nestor A, Plaut DC, Behrmann M (2011): Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proc Natl Acad Sci USA* 108:9998–10003.
- Oldfield RC (1971): The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* 9:97–113.
- Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011): Voice cells in the primate temporal lobe. *Curr Biol* 21:1408–1415.
- Pitcher D, Walsh V, Yovel G, Duchaine B (2007): TMS evidence for the involvement of the right occipital face area in early face processing. *Curr Biol* 17:1568–1573.
- Pitcher D, Walsh V, Duchaine B (2011): The role of the occipital face area in the cortical face perception network. *Exp Brain Res* 209:481–493.
- Pourtois G, Schwartz S, Seghier ML, Lazeyras F, Vuilleumier P (2005): Portraits or people? Distinct representations of face identity in the human visual cortex. *J Cogn Neurosci* 17:1043–1057.
- Proops L, McComb K, Reby D (2009): Cross-modal individual recognition in domestic horses (*Equus caballus*). *Proc Natl Acad Sci USA* 106:947–951.
- Rajimehr R, Young JC, Tootell RB (2009): An anterior temporal face patch in human cortex, predicted by macaque maps. *Proc Natl Acad Sci USA* 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.
- Rossion B (2008): Constraining the cortical face network by neuroimaging studies of acquired prosopagnosia. *Neuroimage* 40: 423–426.
- Rossion B, Caldara R, Seghier M, Schuller AM, Lazeyras F, Mayer E (2003): A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain* 126(Pt 11):2381–2395.
- 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.
- Rotshtein P, Geng JJ, Driver J, Dolan RJ (2007a): Role of features and second-order spatial relations in face discrimination, face recognition, and individual face skills: Behavioral and functional magnetic resonance imaging data. *J Cogn Neurosci* 19: 1435–1452.
- Rotshtein P, Vuilleumier P, Winston J, Driver J, Dolan R (2007b): Distinct and convergent visual processing of high and low spatial frequency information in faces. *Cereb Cortex* 17:2713–2724.
- Sai FZ (2005): The role of the mother’s voice in developing mother’s face preference: Evidence for intermodal perception at birth. *Infant Child Dev* 14:29–50.
- Schweinberger SR, Herholz A, Sommer W (1997a): Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *J Speech Lang Hear Res* 40:453–463.
- Schweinberger SR, Herholz A, Stief V (1997b): Auditory long-term memory: Repetition priming of voice recognition. *Q J Exp Psychol A* 50:498–517.
- Schweinberger SR, Kloth N, Robertson DMC (2011): Hearing facial identities: Brain correlates of face-voice integration in person identification. *Cortex* 47:1026–1037.
- Sheffert SM, Olson E (2004): Audiovisual speech facilitates voice learning. *Percept Psychophys* 66:352–362.
- Sidtis D, Kreiman J (2012): In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integr Psychol Behav Sci* 46:146–159.

- Sliwa J, Duhamel JR, Pascalis O, Wirth S (2011): Spontaneous voice-face identity matching by rhesus monkeys for familiar conspecifics and humans. *Proc Natl Acad Sci USA* 108:1735–1740.
- Smith PL, Ratcliff R (2004): Psychology and neurobiology of simple decisions. *Trends Neurosci* 27:161–168.
- Steeves JKE, Culham JC, Duchaine BC, Pratesi CC, Valyear KF, Schindler I, Humphrey GK, Milner AD, Goodale MA (2006): The fusiform face area is not sufficient for face recognition: Evidence from a patient with dense prosopagnosia and no occipital face area. *Neuropsychologia* 44:594–609.
- Steiger JH (1980): Tests for comparing elements of a correlation matrix. *Psychol Bull* 87:245–251.
- Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008): Neural repetition suppression reflects fulfilled perceptual expectations. *Nat Neurosci* 11:1004–1006.
- Tanji K, Iwasaki M, Nakasato N, Suzuki K (2012): Face specific broadband electrocorticographic spectral power change in the rhinal cortex. *Neurosci Lett* 515:66–70.
- 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.
- Todorovic A, de Lange FP (2012): Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *J Neurosci* 32:13389–13395.
- Tsao DY, Livingstone MS (2008): Mechanisms of face perception. *Annu Rev Neurosci* 31:411–437.
- Von Der Heide RJ, Skipper LM, Olson IR (2013): Anterior temporal face patches: A meta-analysis and empirical study. *Front Hum Neurosci* 7:17.
- 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, 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, 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, Dogan O, Gruter M, Giraud AL, Kell CA, Gruter T, Kleinschmidt A, Kiebel SJ (2008): Simulation of talking faces in the human brain improves auditory speech recognition. *Proc Natl Acad Sci USA* 105:6747–6752.
- Xu X, Yue X, Lescroart MD, Biederman I, Kim JG (2009): Adaptation in the fusiform face area (FFA): Image or person? *Vision Res* 49:2800–2807.
- Zaske R, Schweinberger SR, Kawahara H (2010): Voice aftereffects of adaptation to speaker identity. *Hear Res* 268:38–45.